Data Min Knowl Disc
DOI 10.1007/s10618-012-0295-5

Training and assessing classification rules with imbalanced data

Giovanna Menardi · Nicola Torelli

Received: 13 September 2010 / Accepted: 12 October 2012

Abstract The problem of modeling binary responses by using cross-sectional data has been addressed with a number of satisfying solutions that draw on both parametric and nonparametric methods. However, there exist many real situations where one of the two responses (usually the most interesting for the analysis) is rare. It has been largely reported that this class imbalance heavily compromises the process of learning, because the model tends to focus on the prevalent class and to ignore the rare events. However, not only the estimation of the classification model is affected by a skewed distribution of the classes, but also the evaluation of its accuracy is jeopardized, because the scarcity of data leads to poor estimates of the model's accuracy.In this work, the effects of class imbalance on model training and model assessing are discussed. Moreover, a unified and systematic framework for dealing with the problem of imbalanced classification is proposed, based on a smoothed bootstrap re-sampling technique. The proposed technique is founded on a sound theoretical basis and an extensive empirical study shows that it outperforms the main other remedies to face imbalanced learning problems.

Keywords Accuracy · Binary classification · Bootstrap · Kernel density estimation · Imbalanced learning

Mathematical Subject Classifications (2000): 68T10 · 62G07 · 62G09

Responsible editor: Chih-Jen Lin.

G. Menardi (✉)
Dipartimento di Scienze Statistiche, Università degli Studi di Padova, via C. Battisti, 241, Padova, Italy
e-mail: menardi@stat.unipd.it

N. Torelli
Dipartimento di Scienze Economiche, Aziendali, Matematiche e statistiche "Bruno de Finetti",
Università degli Studi di Trieste, Trieste, Italy

Published online: 30 October 2012

1 Introduction

Classification of new objects, based on the observation of similar instances, is one of the typical tasks in the field of data mining. Here, each object may be denoted by a pair $(\mathbf{x}, y)$ where $\mathbf{x}$ represents a set of measured characteristics, supposed to have some influence on the class label y. In a general framework, $\mathbf{x}$ is defined in a $d-$dimensional space $\mathcal{X}$ being the product set between some continuous, discrete and categorical domains, and the response variable y takes values in the categorical domain $\mathcal{Y} = \{\mathcal{Y}_1, \ldots, \mathcal{Y}_s\}$.

When dealing with a classification task, a sample $T_n = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ of such pairs (the so-called training set) is observed on n individuals or objects and used to build a rule $\mathbf{R}_{T_n} : \mathcal{X} \mapsto \mathcal{Y}$ that allows for the future prediction of the response variable y, based on the observation of $\mathbf{x}$ only.

From a statistical point of view, T_n is usually considered as a collection of i.i.d. realizations from an unknown probability distribution F on $\mathcal{X} \times \mathcal{Y}$.

The rule $\mathbf{R}_{T_n}$ produces a partition of $\mathcal{X}$ in subspaces, each of them associated with a label class $\mathcal{Y}_j$ of $\mathcal{Y}$ and such that:

$$\frac{P(\mathcal{Y}_j|\mathbf{x})}{P(\mathcal{Y}_k|\mathbf{x})} > t, \quad \forall k \neq j, \tag{1}$$

where $P(\mathcal{Y}_j|\mathbf{x})$ is the estimated conditional probability of belonging to $\mathcal{Y}_j$ and t is some threshold t, typically set to 1.

Several techniques have been proposed in the literature for dealing with the classification task: from the simpler approaches as classification trees and nearest neighbors, to the traditional discriminant analysis and multinomial models and the more complex support vector machines, neural networks or ensemble techniques. These methods are basically characterized by some implicit or explicit approach to the estimation of the unknown probabilities involved in Eq. 1. For example, the linear discriminant analysis is based on the assumption of Normality of the $\mathbf{x}|\mathcal{Y}_j$, while the classification trees allocate the data points to the different classes according to a nonparametric estimation of the $P(\mathbf{x}|\mathcal{Y}_j)$.

In this paper, we focus on dichotomic responses, conventionally labeled as negative and positive, that is $\mathcal{Y} = \{\mathcal{Y}_0, \mathcal{Y}_1\}$. In particular, we face the problem of building an accurate classifier when one of the two classes (referred as the positive one) is rare. This class imbalance occurs in many real situations and domains, such as finance (identification of fraudulent credit card transactions or defaulter credit applicants,), epidemiology (diagnosis of cancerous cells from radiographies or any rare disease), social sciences (detection of anomalous behaviors) and computer sciences (feature recognition in image data). Some examples are provided by the recent applications of Ström and Koker (2011); Percannella et al. (2011), and (Pavón et al. 2011).

In certain domains (like those just mentioned), the class imbalance is intrinsic to the problem. However, imbalanced data may occur when the data collection process is limited (for economic or privacy reasons), thus giving rise to an artificial or extrinsic imbalance. Class imbalance may further be absolute or relative (occurring when the cardinality of one class is much larger than the cardinality of the other class, but many

negative and positive examples are observed). See He and Garcia (2009) for further details.

It has been widely reported that the class imbalance heavily compromises the process of learning, because the model tends to focus on the prevalent class and to ignore the rare events (see, e.g., Kubat and Matwin 1997; Japkowicz and Stephen 2002).

A massive interest in imbalanced learning has recently grown, and works focusing on this topic have rapidly reported undeniable advances. However, the research community is still pursuing an undisguised and unified approach to the class imbalance, which remains an open problem (see, e.g., the discussion of He and Garcia 2009). The situation to date appears to provide manifold tools, each of them outperforming the existing methods with regard to some aspect, but being outperformed with regard to other aspects. In many cases, it is not clear why one technique should be preferred to the others (Wasikowski and Chen 2010), and only heuristic reasons are given to justify the suggested proposals. Concerning standard problems of supervised learning, Hand (2006) claims that "the improvements attributed to the more advanced and recent developments are small, […] so that the gains reported on theoretical grounds […] do not translate into real advantages in practice". In imbalanced learning, the perception is that the inverse argument applies, because the lack of a theoretical background, supporting the existing remedies, prevents us from understanding the effectiveness of the various methods.

In addition to the issue of model training in imbalanced learning, a further critical aspect concerns the evaluation of the performance of the model. Among the several approaches to assess such performance (efficiency, interpretability …), the most commonly considered, in classification problems, is the accuracy. In the literature about imbalanced learning, however, while the issue of selecting a suitable metric of accuracy has been fully addressed, a critical inherent aspect has been completely ignored:

whatever metric is chosen for measuring the classifier accuracy, the goodness of the estimate of such metric has not been object of investigation. In fact, such estimate turns out to be very poor when the distribution of the classes is skewed.

A simultaneous treatment of the two inseparable problems of model estimation and evaluation in imbalanced learning has not been considered yet. The main purpose of this work is to address such an issue, and in particular, the following contributes are given:

- a discussion about the effects of class imbalance in binary classification is provided, especially from the perspective of evaluating the accuracy of the estimated classifier, and some numerical examples are given to motivate the emerged issues;
- by taking advantage of some existing ideas, it is shown how to obtain a new technique that outclasses the current remedies to the imbalance problem, both from the theoretical and the computational point of view. This technique, referred to as ROSE (random over sampling examples), is based on a smoothed bootstrap form of re-sampling from data and it hinges on a sound theoretical basis supported by the well-known properties of the kernel methods. Furthermore, it mostly outperforms the main competing the state-of-the-art methods for imbalanced learning;
- it is shown that the proposed technique provides a systematic and unified framework for dealing with imbalanced learning because it may be applied to jointly take

into account the effects of class imbalance in model training and model assessing, at least when the distribution of the classes is highly skewed.

Section 2 discusses the effects of a highly skewed distribution of the classes when building and measuring the accuracy of classification rules. In Sect. 3, our technique to deal with imbalanced learning is presented, its properties are enlightened and a comparison to some similar existing remedies is discussed. Section 4 presents some numerical evidence about the use of ROSE. Some final remarks conclude the paper.

2 The effects of class imbalance

In many practical applications of binary classification problems, an extremely imbalanced distribution of the two label classes has been found. In principle, the issue might be tackled by the standard application of any supervised method of classification, such as the ones mentioned in the previous section. However, unless the classes are perfectly separable (Hand and Vinciotti 2003) or the complexity of the problem is low (Japkowicz and Stephen 2002; Batista et al. 2004), neglecting the unbalance leads to heavy consequences, both in model estimation and when the evaluation of the accuracy of the estimated model has to be measured. Providing a complete review of the inherent literature is beyond the scope of this paper (for two excellent reviews, see, e.g. He and Garcia 2009; Sun et al. 2009). However, the current section aims at understanding the main issues that emerge in modeling and assessing the accuracy of imbalanced data.

2.1 Model estimation in the presence of rare classes

Failure of classification methods when the model estimation is based on a skewed training set is a very well-known problem in the literature. What typically happens in that situation is that standard classifiers tend to be overwhelmed by the prevalent class and ignore the rare examples.

It has been largely reported that, whatever standard classification method is chosen, such a failure occurs in non-trivial learning problems. Nonetheless, the reasons for the occurrence of this behaviour are strongly dependent on the choice of the method.

Logistic regression, for instance, known as one of the most traditional parametric methods for binary classification, is not advisable when the classes are imbalanced, because the conditional probabilities of the rare class are underestimated (King and Zeng 2001). The use of logit models in classification with skewed data is also discussed by Cramer (1999); King and Ryan (2002) and, more recently by Oommen et al. (2011).

Also the performance of linear discriminant analysis is compromised when the distribution of the classes is imbalanced. The estimate of the common covariance matrix of the two classes is a weighted mean of the two sample matrices, hence being dominated by the dispersion of the prevalent class. If the assumption of equal covariance matrix does not hold, a substantial bias may ensue. The issue is discussed in Hand and Vinciotti (2003).

Not even the more flexible nonparametric methods are immune to the consequences of a skewed distribution of the classes. Those classifiers are designed to build the classification rule that best fits the data according to the optimization of some objective function. When this function is based on a criterion of global accuracy, the classifier tends to favor classification rules that perform well only on the frequent class (see the next section for further details about using overall accuracy measures in imbalanced learning). This is the case, e.g., for classification trees, discussed in the context of imbalance learning by Chawla (2003) and Cieslak and Chawla (2008) or for support vector machines (Akbani et al. 1999).

As the choice of model complexity in classification trees leads to a trade-off between bad fitting and poor generalization in a skewed-class framework so, the choice of k, when the $k-$ nearest neighbor classifier is used, gives rise to clashing opinions. Kubat and Matwin (1997), for instance, observe that with large samples, the performance of the classifier may improve if k neighbors are used, instead of one. In contrast, Hand and Vinciotti (2003) claim that the probability of correctly classifying an example from the minority class is a decreasing function of k. Anyway, in both works, authors agree that the critical point is that k should be much smaller than the small class, which is often a problem when one class is rare.

An argument that moves away from the ones just mentioned, suggests that the effects of class imbalance are not directly caused by the distribution of the classes, but rather that the imbalance may lead to small disjuncts determining, in turn, a degradation of the classification (Jo and Japkowicz 2004).

Most of the current research on imbalanced classification focuses on proposing solutions to improve the stage of model estimation. The literature is wide, and the remedies are various (see, for a review, Kotsiantis et al. 2006). However, the main contributions can be summarized in solutions at the learning level and solutions at the data level.

1. Solutions at the learning level aim at strengthening the learning process towards the minority class. A first approach consists of making some modification of the classifier in order to compensate for the imbalance. This approach is generally applied to classifiers whose training is based on the optimization of a function related to the overall accuracy. Improvements of the learning ability are then achieved by using alternative functions that are independent of the distribution of the classes. This is the route followed, for instance, by Riddle et al. (1994) and Cieslak and Chawla (2008) when using decision trees, Veropoulos et al. (1999) and Batuwita and Palade (2010) when support vector machines are used and Barandela et al. (2003) that consider $k-$nearest neighbor classifiers.
   Other remedies, focusing on modification of the learning process, give different misclassification costs to the training data in order to force the classification rule to focus more on the positive examples. This approach is usually followed when the skew distribution of the classes is associated with different misclassification costs (typically, the cost of misclassifying a rare example is higher than the corresponding cost for a negative example). In these cases, a classification rule that minimizes the expected misclassification cost is trained. Most learning methods may be easily modified in order to take into account the cost of misclassification. In decision trees, for instance, a cost function may be introduced in the splitting

criterion, as pointed out by Breiman et al. (1984) and Ting (2002). Similar ideas are discussed by Kukar and Kononenko (1998) in the context of neural networks and by Lin et al. (2002) who develop cost-sensitive support vector machines. One problem with this approach is that specific cost information is usually not available.

Alternatively, a general technique for introducing different propensities toward misclassification errors consists of moving the classification threshold in Eq. 1 toward the less expensive class so that examples of the minority class become harder to be misclassified. It is easy to show that, given that c_j is the cost of misclassifying a class $\mathcal{Y}_j$ object, the minimum loss is achieved by assigning an observation to the class $\mathcal{Y}_1$ if $c_0 P(\mathcal{Y}_0|\mathbf{x}) > c_1 P(\mathcal{Y}_1|\mathbf{x})$, that is, if the classification threshold is set to c_0/c_1. Examples of this approach can be found in Eitrich et al. (2007) as well as in Zhou and Liu (2006).

Remedies at the learning level also include the use of combinations of classifiers, by following logics typical of boosting, bagging or random forests. These methods are, by construction, computationally demanding. Some references are Liu et al. (2006); Thomas et al. (2006); Sun et al. (2007); Khoshgoftaar et al. (2007, 2011); Fernandez et al. (2012).

A further approach that has been recently shown to lead an improvement of classification in the presence of imbalanced classes makes use of evolutionary algorithms for generalized instances selection (García et al. 2011).

The learning approaches have often resulted in effectively limiting the consequences of the class imbalance when training the classifier, but they have the disadvantage of being algorithm-specific, while data sets presenting different characteristics are better treated by different classification methods.

2. Solutions at the data level for dealing with imbalanced classes focus on altering the class distribution to get a more balanced sample.

   Remedies following this approach include various techniques to sample the data, as random oversampling with replacement the rare class and random undersampling (without replacement) the prevalent class. Oversampling, in its simplest form, duplicates examples of the minority class, while undersampling removes some data from the frequent class. The characteristics of both these sampling techniques have been widely studied (Japkowicz and Stephen 2002; Estabrooks et al. 2004) and considered in various applied works (see, e.g., Burez and Van den Poel 2009; Mazurowski 2008) also in conjunction with other remedies (Wu and Zhou 2009). Moreover, they are usually suggested by some commercial data mining software (e.g., SAS Enterprise Miner) as the main remedy to be adopted. Slight modifications of the mentioned techniques are directed oversampling or undersampling (where the choice of examples to duplicate or, respectively, remove is informed instead of random), or combinations of these techniques (Kubat and Matwin 1997; Yen and Lee 2006).

   Both oversampling and undersampling decrease the overall level of class imbalance, thereby improving the overall accuracy of the classifier. The reason is that altering the class distribution corresponds to impose non-uniform misclassification costs. This equivalence is well-known and was first formally elucidated in Breiman et al. (1984).

Both undersampling and oversampling have known drawbacks (McCarthy et al. 2005). Undersampling may discard potentially useful data, thus reducing the sample size, while oversampling may increase the likelihood of overfitting, since it is bound to produce ties in the sample, especially as the sampling rate increases. Moreover, the augmented sample increases the computational effort of the learning process.

Increasing attention has been recently paid to the novel strategy of generating new artificial examples that are "similar" in some sense to the observations belonging to the minority class.

In Lee (1999), for instance, a fixed number of replicates of each rare event is created, by adding some normal noise to the trained observations. The $P(\mathcal{Y}_j|\mathbf{x})$ are then estimated by the application of some standard binary classifier and possibly averaged across a number of iterations (Lee 2000). Chawla et al. (2002) propose a method called synthetic minority oversampling technique (SMOTE). For each rare training observation, new examples are generated by randomly choosing points that lie on the line connecting the rare observation to one of its nearest neighbors in the feature space. The same idea is then extended to an improved boosting algorithm for dealing with rare classes. Similarly, boosting is combined with data generation in Guo and Viktor (2004) and Mease et al. (2007).

Generation of new artificial data that have not been previously observed reduces the risk of overfitting and improves the ability of generalization compromised by the oversampling methods. For this reason, this is also the approach followed in this paper.

2.2 Model evaluation in the presence of rare classes

When a classification task is performed, evaluating the accuracy of the classifier plays a role that is at least as important as the model estimation, especially in a class imbalance framework. Indeed, both the choice of the best classification rule among alternative ones, and the extent to which a classification rule may be operatively applied to real-world problems for labeling new unobserved examples, depend on our ability to measure the classification accuracy.

Although the literature about model assessment in a class imbalance framework has been fast developing recently, the issue has not yet received as much attention as the one focusing on the stage of model training. In fact, even if an effective classification rule was trained on the data, the class imbalance would still lead to non-negligible consequences when evaluating the model accuracy. Two problems arise in model assessment in the presence of imbalanced classes, concerning the choice of the evaluation measure and the estimate of such a measure of accuracy.

1. It has been largely emphasized (He and Garcia 2009; Weiss and Provost 2001; Weiss 2004) that the use of common performance measures, such as the error rate, may yield to misleading results because they strongly depend on the class distribution. For instance, in a problem where the rare class is represented in only 1 % of the data, the naive strategy of allocating each example to the prevalent class would achieve a good level of accuracy, presenting an overall error rate equal

to 1 %. However, it is clear that such a classification rule is completely useless. Hence, the choice of the evaluation measure has to be addressed toward some class-independent quantities.

To this aim, more appropriate performance measures may be derived from the observation of the confusion matrix, which compares the predicted labels to the true labels. In order to provide comprehensive assessments of imbalanced learning problems, the most frequently adopted performance measures are based on different propensity towards false negatives (FN) and false positives (FP). Precision, for instance, computes the fraction of examples classified as positive that are truly positive, while recall measures the fraction of correctly labeled positive examples. Precision is sensitive to the distribution of the classes whereas recall is not. However, recall provides no insight as to how many examples are incorrectly labeled as positive, so the two measures have to be used jointly. Alternatively, precision and recall may be combined into their geometric mean or into a more elaborate summarizing function called the F measure. Similarly, the G mean computes the geometric mean of the accuracies of the two classes.

One of the most frequently used tools for evaluating the accuracy of a classifier in the presence of imbalanced classes is the receiver operating characteristics (ROC) curve. As the classification threshold varies, the predicted label is assigned to the examples and the confusion matrix represented. The true positive rate (sensitivity of the classifier) is then plotted versus the false positive rate (1—specificity of the classifier) for each considered value of the classification threshold. The classifier performs as better the steeper the ROC curve becomes, that is, the larger the area underlying the curve (AUC) is. A completely random guess would give rise to a ROC curve lying along the diagonal line from the bottom left to the top right corners, whereas a perfect classifier would yield a point in the upper left corner of the ROC space, representing 100 % sensitivity (all true positives are found) and 100 % specificity (no false positives are found). ROC curves can help to compare different trade-offs arising from the use of distinct classifiers. However, they do not take into account different misclassification cost and class distributions.

Similarly, precision-recall curves may be adopted for assessing the classification accuracy (Davis and Goadrich 2006) and cost curves feature the ability to compare the performance of a classifier over a range of misclassification costs and class distributions (Drummond and Holte 2006). For a complete review about the evaluation metrics in a class imbalance framework, see, for instance, He and Garcia (2009).

2. Although the most frequently adopted evaluation metrics share some drawbacks, the research focusing on this issue has been very fruitful and several advances have been made.

   In fact, the evaluation of the accuracy of a classifier in imbalanced learning is subject to a more serious problem than the choice of an adequate error metric. This problem concerns the estimate of such accuracy and, as far as we know, it has been completely neglected by the literature.

   In learning problems, one is interested in measuring the accuracy of a classifier by its ability to assign a previously unseen example $(\mathbf{x}_0, y_0)$ to the correct

class. Given a classification rule $\mathbf{R}_{T_n}$, based on a training set $\mathbf{T}_n$, a $0-1$ loss function $L\left((\mathbf{x}_0, y_0), \mathbf{T}_n, \mathbf{R}_{T_n}\right)$ is typically used to define the *true* or *conditional error*:

$$Err = E_{F(\mathbf{x}_0, y_0)}\left[L\left((\mathbf{x}_0, y_0), T_n, \mathbf{R}_{T_n}\right)\right] \tag{2}$$

where the expectation is taken with respect to the probability distribution F of $(\mathbf{x}_0, y_0)$ and T_n. Clearly, the expression of the error measure changes if the accuracy is measured by using alternative performance criteria such as the precision, recall or the AUC. However, the key matter is that since F is unknown, an estimate of Eq. 2 has to be considered. Popular error estimators are the apparent error (also called resubstitution method) and the holdout method. The former measures the accuracy of the classifier on the training set, while the latter consists of dividing all the available data into two disjoint sets, used for training the classifier and testing its accuracy. Other estimators are based on bootstrap or cross-validation ideas. For a review, see, for example, Schiavo and Hand (2000). As far as the research community continues to develop and apply more advanced performance criteria for dealing with imbalanced classes, it seems that the possible consequences of neglecting the quality of such criteria have not been considered. In fact, poor estimates of the classifier's performance may lead to misleading conclusions about the quality of the classifier, and proposing more and more sophisticated learning methods becomes a wild-goose chase if we are not able to evaluate their accuracy. In most of the literature about classification in the presence of rare classes, the empirical analysis consists of estimating the classifier over a training set and assessing its accuracy on a test set. However, in real data problems, there are not enough examples from the rare class for both training and testing the classifier, and the scarcity of data conducts to high variance estimates of the error rate, especially for the rare class.

The issue will be clarified in Sect. 4.2 by giving some evidence about this fact. While it is clear and widely acknowledged that validating the classifier on the same data used in the training stage may lead to too optimistic conclusions about the accuracy of the estimated model, it turns out that, in strongly imbalanced frameworks, also the practice of validating results on a test sample may be not appropriate because of the scarcity of rare examples. Then, alternative estimators of the chosen accuracy metrics have to be considered, as the next Section will enlighten.

3 Random oversampling examples

In the previous section, it has been outlined that the performance of classification models is comprehensively compromised by a skewed distribution of the classes, but, even worse, poor-quality estimates of the chosen accuracy measure may preclude understanding the limits of the learning process. It stands to reason that a new perspective for approaching the issue of class imbalance should be considered, and the problems of building an accurate classifier and assessing its performance should not be dealt with separately.

The main contribution of this work consists of providing a unified and systematic framework for simultaneously dealing with these two inseparable problems. We follow the traditional approach of balancing the distribution of the classes, both because of the flexibility of this approach in supporting the application of any classification method and because it allows a natural joint treatment of the issues emerging from the estimation and assessment of the classifier. The proposed solution may be referred to as ROSE, and it is based on the generation of new artificial data from the classes, according to a smoothed bootstrap approach (see, for example, Efron and Tibshirani 1993).

We focus on $\mathcal{X}$ domains included in R^d, and $P(\mathbf{x}) = f(\mathbf{x})$ is a probability density function on $\mathcal{X}$. Let $n_j < n$ be the size of $\mathcal{Y}_j$, $j = 0, 1$. The ROSE procedure to generate one new artificial example consists of the following steps:

1. select $y = \mathcal{Y}_j \in \mathcal{Y}$ with probability $\frac{1}{2}$
2. select $(\mathbf{x}_i, y_i)$ in $\mathbf{T}_n$ such that $y_i = y$ with probability $p_i = \frac{1}{n_j}$
3. sample $\mathbf{x}$ from $K_{\mathbf{H}_j}(\cdot, \mathbf{x}_i)$, with $K_{\mathbf{H}_j}$ a probability distribution centered at $\mathbf{x}_i$ and depending on a matrix $\mathbf{H}_j$ of scale parameters.

Essentially, we draw from the training set an observation belonging to one of the two classes (chosen by giving the same probability to $\mathcal{Y}_0$ and $\mathcal{Y}_1$) and generate a new example in its neighborhood, where the width of the neighborhood is determined by $\mathbf{H}_j$. Usually, $K_{\mathbf{H}_j}$ is chosen in the set of the unimodal, symmetric distributions. It is worthwhile to note that, once that a label class has been selected,

$$\begin{aligned}\hat{f}(\mathbf{x}|y = \mathcal{Y}_j) &= \sum_{i=1}^{n_j} p_i \, Pr(\mathbf{x}|\mathbf{x}_i) \\ &= \sum_{i=1}^{n_j} \frac{1}{n_j} Pr(\mathbf{x}|\mathbf{x}_i) \\ &= \sum_{i=1}^{n_j} \frac{1}{n_j} K_{\mathbf{H}_j}(\mathbf{x} - \mathbf{x}_i).\end{aligned}$$

It follows that the generation of new examples from the class $\mathcal{Y}_j$, according to ROSE, corresponds to the generation of data from the kernel density estimate of $f(\mathbf{x}|\mathcal{Y}_j)$, with smoothing matrix $\mathbf{H}_j$. A discussion about how to set the $\mathbf{H}_j$ matrices is given in the next Section.

The repeated implementation of steps 1–3 allows for the creation of a new synthetic training set $\mathbf{T}_m^*$, with size m where approximately the same number of examples belong to the two classes. The size m may be set to the original training set size n or chosen in any way. ROSE combines techniques of oversampling and undersampling by generating an augmented sample of data (especially belonging to the rare class) thus helping the classifier in estimating a more accurate classification rule, because the same attention will be addressed to both the classes.

However, the synthetic generation of new examples allows for strengthening the process of learning as well as estimating the distribution of the chosen measure of

accuracy. Operatively, the artificial training set $\mathbf{T}_m^*$ may be used to estimate the classification model, and the originally observed data remain free of being used for testing the classifier. Alternatively, cross-validation or smoothed bootstrap methods can be used. It is worth noting that creating new artificial examples from an estimate of the conditional densities of the two classes allows for overcoming the limits of both the apparent error (providing too optimistic evaluations of the classifier performance) and the holdout method (non-advisable in extreme imbalanced learning because the scarcity of rare class data prevents their use in both estimating and testing the model).

3.1 Selecting the smoothing matrix $\mathbf{H}_j$

The operational use of ROSE requires a prior specification of the $\mathbf{H}_j$ matrices. In principle, the issue is quite critical, since different choices of the smoothing matrices lead to larger or smaller $K_{\mathbf{H}_j}$, namely larger or smaller neighborhoods of the observations from which the synthetic examples are generated. However, as ROSE draws artificial data from the kernel density estimates of the conditional densities of the observations, the choice of the $\mathbf{H}_j$ may be addressed by the large specialized literature on methods to choose the smoothing parameters (some excellent reviews are, for instance, Silverman 1986; Bowman and Azzalini 1997).

The idea at the basis of the these methods is to minimize an optimality criterion as for example the asymptotic mean integrated squared error (AMISE). The AMISE still depends on some unknown quantities and an approximation has to be used. Examples are bootstrap, cross-validation or plug-in estimates of the AMISE. A detailed description of these methods (which are directly available on the main statistical softwares) is beyond the scopes of this work, and the reader could refer to the specialized literature. In the following, however, the selection rule that has been used in this work is described, with the aim of providing the reader with an operational tool.

Among the many possible alternatives, we consider Gaussian Kernels with diagonal smoothing matrices $\mathbf{H}_j = diag(h_1^{(j)}, \ldots, h_d^{(j)})$ and we minimize the AMISE under the assumption that the true conditional densities underlying the data follow a Normal distribution. This leads to (see, e.g., Bowman and Azzalini 1997, p. 32):

$$h_q^{(j)} = \left(\frac{4}{(d+2)n}\right)^{1/(d+4)} \hat{\sigma}_q^{(j)} \quad (q = 1, \ldots, d;\ j = 1, 2). \tag{3}$$

Here, $\hat{\sigma}_q^{(j)}$ is the sample estimate of the standard deviation of the $q-th$ dimension of the observations belonging to the class $\mathcal{Y}_j$.

While this strategy may appear naive, it has resulted in producing good results in virtually all cases on which we have tested the method. Indeed, we are not interested to build an accurate estimate of the true conditional densities, but just to generate new examples in a reasonable neighborhood of the observations. Moreover, ROSE generates data separately from each class. Hence, our choice looks well advised because, even if the $f(\mathbf{x}|y = \mathcal{Y}_j)$ may not be Normal, we are confident that they should be at least unimodal, because each of them describes one class only.

3.2 Discussion

The idea of oversampling the rare class (or combining minority oversampling and majority undersampling), to provide for the imbalance of the classes, has been already developed by several authors. ROSE generalizes the standard technique of oversampling with replacement the rare examples by allowing the generation of some clones of the observed data, without producing ties. As mentioned by Chawla et al. (2002), generating synthetic examples, instead of simply increasing the multiplicity of the rare events, has the desirable effect of causing the classifier to identify larger decision regions associated to the minority class.

On the other hand, ROSE includes oversampling with replacement as a special case, occurring when the elements of $\mathbf{H}_j$ tend to zero.

Indeed, also the idea of generating artificial examples similar, in some way, to the observed data has been already discussed in several works. However, unlike those works, ROSE has some features which make its use preferable:

- While it is clear that the necessity to break ties (when changing the multiplicities due to the oversampling) motivates the choice of generating new artificial examples, the works that use this approach do not clarify why such data generation should be performed according to the proposed solutions, and only heuristic reasons are given to justify the choice. In contrast, ROSE is founded on a sound theoretical basis supported by the well-known properties of the kernel methods. ROSE draws synthetic examples from an estimate of the (conditional) density underlying the data, thus providing confidence that the distribution of the data into the classes has not changed since the balancement has been performed.
- In order to perform the data generation, it is unavoidable that all the existing techniques leave one or more parameters to be user-defined. Definition of such parameters either requires some high computational effort or is based on some vague mechanism. In Chawla et al. (2002), for instance, the number of nearest neighbors to be considered for each rare example is an input parameter, while in Lee (1999, 2000), the generation of new events depends on a scale parameter whose optimum value is determined according to a computationally intensive iterative procedure. Similarly, ROSE requires that the $\mathbf{H}_j$ matrices are defined beforehand for each class. However, as discussed in Sect. 3.1, the link between ROSE and the kernel methods allows us to consider each $\mathbf{H}_j$ as a smoothing matrix. Hence, it is possible to take advantage of the huge literature about kernel density estimation and choose the $\mathbf{H}_j$ as the solutions of one of the several methods of bandwidth selection.
- As previously observed, generating artificial data allows for exploiting the original observations for testing the accuracy of the estimated model. In this way, the necessity of a preliminary splitting of the data into a training set and a validation set, which entails a loss of information useful to the stage of learning, is avoided. However, none of the mentioned works take advantage of this potentiality.

Special attention should be paid to comparing ROSE to the solutions proposed by Lee (1999, 2000), which, at first glance, present many similarities. The author suggests creating new occurrences of the rare cases by adding some normal noise to the

observed events. Hence, when a Gaussian kernel is chosen in applying ROSE, the mechanism for generating one new rare example is the same. However, it is worthwhile to note some practical differences, also affecting the theoretical interpretation of the two methods, which aid considering ROSE as an improved generalization of the contribution proposed by Lee.

- While Lee increases the occurrence of the rare cases only and leaves the prevalent examples unchanged, in ROSE, the data generation involves both the minority and the majority class. Hence, the synthetic and the observed training sets do not even partially overlap, thus reducing the risk of overfitting and giving the opportunity of using the observations for model testing.
- In work by Lee, the occurrence of rare examples is exactly multiplied by a predetermined constant. The value of such a constant is user-defined but results from a simulation study suggest doubling of the cardinality of the minority class. ROSE creates an artificial sample where data belonging to the two classes have the same probability of occurrence, thus giving rise to a balanced sample. While in principle, our choice should allow for dealing with even extremely imbalanced data, doubling the size of the rare class may help the learning process only in moderately imbalanced situations.
- In work by Lee, all the minority examples give rise to a fixed number of noise replications. On the other hand, in ROSE, a random selection guides the choice of the observations from which the artificial examples are created (within each class, the observations are given the same probability of selection), thus making possible the interpretation of the strategy of data generation as the selection of a smoothed bootstrap sample (except that the new artificial classes do not have the same size as the original ones).
- Lee draws each noisy replicate from a normal distribution centered at an observed minority class example and with diagonal covariance matrix proportional to the vector of sample variances of the explanatory variables $\mathbf{x}$. This procedure allows for a better estimate of the covariance matrix (since it is based on a larger sample). However, the choice corresponds to the assumption that the two classes have a common covariance matrix, which is not, in general, true. In ROSE, the smoothing matrices are evaluated by using the data belonging to the two classes separately. Moreover, it should be argued that using a diagonal covariance matrix leads to the generation of the new data from a spherical distribution and, hence, the new artificial sample will not follow the direction of the original data.
- Also, the choice of the kernel is not indifferent when new data have to be generated. Although the literature concerning kernel density estimation agree that the selection of the kernel function is not critical, there are situations in which the Gaussian distribution is not advisable (for instance, when the data have a bounded support, or when reduction of bias is of interest, as mentioned by Silverman 1986).
- An improved version of the technique of Lee (1999) is described in Lee (2000), aimed at reducing the variance of the estimated conditional probabilities of the data and show even more substantial differences from ROSE. For a given data set, several noisy training samples are independently generated and a classifier is trained on them. Afterward, the estimated conditional probabilities obtained by

each classifier are averaged across the samples. Combining several versions of the same classifier is known to improve the performance of a single model, although the computational complexity increases. However, in extremely imbalanced learning, it is not clear if the gain in accuracy is worth the increased computational effort (see the results from the simulation study below). Moreover, even when more classifiers are combined to get an improved estimate of the conditional probabilities, generating new samples through ROSE is more attractive than using the procedure proposed by Lee. In fact, repeatedly bootstrapping the data from the two classes with ROSE, prior to estimating the model, has the beneficial interpretation of building a bagging classifier (Breiman 1996).

4 Empirical analysis

The current section aims to understand if the good theoretical properties of ROSE are associated to good performance in practice, when ROSE is used in imbalanced frameworks. Some empirical analysis have been conducted with the following goals:

1. evaluating the ability of ROSE to strengthen the learning process in an imbalanced classification problem and comparing its behavior with other remedies to deal with skewed distributions of the classes;
2. analyzing the opportunity to exploit ROSE to assess the classification;
3. showing how to use ROSE operationally in an imbalanced framework;

Analysis have been accomplished on 2 simulated data structures and 20 real data sets, varying extensively in the sample size and the class proportions.

In the first simulated set of examples, data have been generated from a mixture of bivariate Normal distributions with fixed means and covariance matrices but varying mixing proportions in order to manage different levels of imbalance. The second simulated structure of data is defined in R^{10}. The rare class may be described as a depleted semi-hypersphere filled with the prevalent class, which is normally distributed and has elliptical contours. Again, different class proportions have been considered. Further details about the two simulated examples are given in Table 1.

Concerning the real data applications, a description of the data sets is summarized in Table 2. Whenever the response variable was not binary, the classes have been selected or suitably aggregated. As most of the considered data sets did not generally present characteristics of extreme imbalance, examples from the minority group have been

Table 1 Simulated data examples

Mixture of normal data	$(\mathbf{x}, y)$ s. t. $\begin{cases} \mathbf{x} \sim \mathbf{N}_2\,(\mathbf{0}_2, \mathbf{I}_2) & \text{if } y = 0 \\ \mathbf{x} \sim \mathbf{N}_2\,\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & -0.5\\ -0.5 & 1\end{pmatrix}\right) & \text{if } y = 1 \end{cases}$	
Filled semi—hypersphere data	$(\mathbf{x}, y)$ s. t. $\begin{cases} \mathbf{x} \sim \mathbf{N}_{10}\,(\mathbf{0}_{10}, (0.25, \mathbf{0}_9)\mathbf{I}_{10}) & \text{if } y = 0 \\ \mathbf{x} \sim \mathbf{N}_{10}\,(\mathbf{0}_{10}, \mathbf{I}_{10}) \cap \|\mathbf{x}\| < 4 \cap x_1 \leq 0 & \text{if } y = 1 \end{cases}$	

Notation: $\mathbf{0}_d$ and $\mathbf{I}_d$ denotes a vector of zeros and, respectively, the identity matrix in R^d; $\|\mathbf{x}\|$ is the norm of vector $\mathbf{x}$, being $\mathbf{x} = (x_1, \ldots, x_d)$.

Table 2 Real data set description: data name, source, selected binary response, number of quantitative variables, data set size (after removing not considered classes and missing values), and proportion of selected examples from the two classes

Name	Source	Classes (1/0)	♯ Variables	♯ Cases	Imbalance
Wine quality	UCI	$\leq 4/ > 4$ Wine quality	11	4898	0.4–99.6
Forest cover	UCI	Ponderosa pine/ Cottonwood-willow	9	38501	0.5–99.5
Infocamere	on request	Defaulter/non-defaulter firms	27	11199	0.7–99.3
Abalone	UCI	$>20/\leq 20$ Rings	7	4177	0.8–99.2
Hypothyroid	UCI	Positive/negative to hypothyroidism	25	2398	1–99
Adult	UCI	>50K \$/$\leq$ 50K \$ Income	6	48842	2–98
Phoneme	ELENA	Oral/nasal sounds	5	5404	2.5–97.5
Breast cancer	UCI	Malignant/benign cytology	30	569	3–97
Cardiotocography	UCI	Suspects/normal cardiotocograms	23	1950	3.5–96.5
Transfusion	UCI	Blood donated yes/no	4	748	4–96
Glass	UCI	Windows float proc yes/no	9	214	4.5–95.5
Pima indians	UCI	Positive/negative diabetes test	8	768	5–95
Cylinder bands	UCI	Band/no band	20	365	6–94
Vehicle silhouettes	UCI	Opel/saab	18	429	7–93
Image segmentation	UCI	Window image yes/no	19	2310	8–92
Spectf	UCI	Diagnosis normal/abnormal	44	267	9–91
Vertebral column	UCI	Normal/abnormal	6	310	10–90
Parkinsons	UCI	Disease yes/no	22	195	12.5–87.5
Concrete compressive strength	UCI	$\geq 50 /< 50$ Strength	8	1030	15–85
Credit screening	UCI	Granted credit yes/no	6	653	20–80

randomly selected in order to cover different levels of extreme skewness. To perform the analysis, the quantitative variables only have been selected from the original data sets and missing values have been excluded. All but two examined data sets are publicly available at the UCI machine learning repository (Asuncion and Newman 2007).

Sections 4.1, 4.2 and, respectively, 4.3 report how the designed tasks have been operatively performed and show the results.

4.1 Model estimation by using ROSE

In order to evaluate the effectiveness of ROSE as a remedy to strengthen the estimation step in classification problems with skewed data, the procedure has been applied to the described simulated and real data sets. The learning methods adopted in the empirical study are nonparametric classification trees and logit models. The area under the ROC curve (AUC) has been chosen as an evaluation metric.

Instead of exploiting the opportunity offered by ROSE to use the artificial sample to train the classification rule and the original data to test it, the classifiers have been

Table 3 Simulation design: π is the proportion of rare examples in the training set (here, a fixed number of observations has been drawn from each class in order to be sure that the rare class does is not empty); $mult$ is the number of noisy replicates for each rare example in Lee (2000); hence, the training set size is $mult\cdot$ number of minority examples+ number of majority examples. To compare the methods on equal terms the balanced sample generated according to ROSE and SMOTE has the same size

Classifiers	Classification trees, logit model
(Original) Training set size n	250, 1000, 5000
Test set size	250
π	0.1, 0.05, 0.025, 0.01
$mult$	2,5,10
K	5,10,20
number of simulations	100

K is the number of noisy training sets generated for a given training set according to Lee (2000) and the number of bagging iterations when the data are repeatedly bootstrapped according to ROSE

estimated on a training set while their performance has been evaluated on a test set (that is, a hold-out method has been used). The use of simulated data has allowed to vary extensively both the training set size and the proportion of events belonging to the minority class. Details about the simulation design are given in Table 3. Concerning the real data, training sets having half size of the original data have been selected and the remaining data have been used to test the models.

ROSE has been compared to some competitors designed to deal with imbalanced classification problems: undersampling, SMOTE and the regularization practice proposed by Lee (2000). Moreover, a bagging version of ROSE has been tested on the simulated data by repeatedly bootstrapping the two classes. The estimation of a standard classifier without using any remedy for dealing with the imbalance has been considered as a benchmark.

Undersampling has been implemented by randomly selecting a subsample of the prevalent class from the training set so that two balanced classes might be obtained. The application of SMOTE has been performed by choosing 5 nearest neighbors, as suggested by the authors in their paper. Whenever less than 6 rare examples were observed in the data, the whole minority class has been used as a set of nearest neighbors. In applying the procedure proposed by Lee (2000), the scale parameter has been set to the optimum value resulting from the simulation study he carried out, while the smoothing matrices in ROSE have been selected according to expression (3). Since the application of the regularization method proposed by Lee leads to an artificial training set with a different size from the original data (because all the majority examples are used and the cardinality of the minority class is multiplied by a predetermined constant), balanced synthetic samples of the same size have been generated according to ROSE and SMOTE.

In Tables 4, 5, and 6, the results referring to the use of a 8-nodes tree on the simulated sets of data are reported.

No surprising results arise from the application of a standard classification tree without resorting to any remedy for the imbalance: regardless of the original sample size n, the accuracy of the classifier decreases with the proportion of rare examples.

Table 4 AUC obtained when training an 8-node classification tree without using any remedy for imbalanced data (left) and by undersampling (right)

π	Imbalanced data Mixture of normal data			Filled semi—hypersphere data			Undersampling Mixture of Normal data			Filled semi—hypersphere data		
	n			n			n			n		
	250	1000	5000	250	1000	5000	250	1000	5000	250	1000	5000
0.10	0.67	0.73	0.62	0.69	0.79	0.80	0.67	0.76	0.80	0.66	0.79	0.78
0.05	0.52	0.61	0.51	0.60	0.76	0.64	0.62	0.70	0.76	0.63	0.71	0.81
0.025	0.50	0.55	0.51	0.54	0.71	0.69	0.56	0.63	0.76	0.61	0.69	0.83
0.01	0.50	0.50	0.50	0.53	0.65	0.65	0.51	0.63	0.72	0.63	0.68	0.72

Cf. 3 for notation

Table 5 AUC obtained when training an 8-node classification tree after balancing the sample by ROSE and SMOTE

π	Mixture of normal data						Filled semi—hypersphere data					
	$mult = 2$			$mult = 10$			$mult = 2$			$mult = 10$		
	n			n			n			n		
	250	1000	5000	250	1000	5000	250	1000	5000	250	1000	5000
ROSE												
0.10	0.79	0.78	0.80	0.78	0.79	0.80	0.68	0.80	0.80	0.73	0.80	0.82
0.05	0.81	0.80	0.82	0.80	0.83	0.81	0.67	0.77	0.81	0.68	0.79	0.80
0.025	0.75	0.77	0.80	0.77	0.78	0.77	0.65	0.78	0.82	0.65	0.76	0.79
0.01	0.72	0.76	0.78	0.72	0.78	0.76	0.64	0.73	0.81	0.65	0.72	0.79
SMOTE												
0.10	0.69	0.75	0.78	0.68	0.77	0.82	0.64	0.82	0.84	0.64	0.81	0.74
0.05	0.61	0.76	0.81	0.64	0.74	0.79	0.61	0.74	0.76	0.59	0.79	0.75
0.025	0.58	0.71	0.75	0.59	0.73	0.77	0.57	0.69	0.81	0.56	0.67	0.74
0.01	0.52	0.69	0.71	0.54	0.71	0.76	0.55	0.65	0.78	0.55	0.65	0.76

Cf. 3 for notation

Moreover, when the minority observations amount to only 1 % of the training set and the sample size is not large, the classifier does not even perform better than a random guess.

When undersampling is used as a remedy to the imbalance, a slightly better classification may be achieved. However, undersampling data has some effectiveness when large samples only are used, because balancing the data determines an extreme reduction of the training set size.

Combining undersampling the majority class and oversampling the minority class by creating synthetic rare examples, as both SMOTE and ROSE do, allow for a remarkable improvement of the accuracy (Table 5). The empirical analysis shows that, in

Table 6 For both the regularization method proposed in Lee (2000) and ROSE, each of the 12 subtables reports the AUC of classification performed on a n-sized sample with proportion of positive examples set to π

	π	Mixture of normal data						Filled semi—hypersphere data					
		$mult = 2$			$mult = 10$			$mult = 2$			$mult = 10$		
		n			n			n			n		
		250	1000	5000	250	1000	5000	250	1000	5000	250	1000	5000
Lee regularization													
$K = 5$	0.10	0.79	0.79	0.78	0.83	0.84	0.81	0.69	0.78	0.81	0.77	0.82	0.82
	0.05	0.72	0.73	0.78	0.85	0.81	0.82	0.64	0.75	0.74	0.72	0.81	0.80
	0.025	0.65	0.79	0.70	0.79	0.85	0.77	0.62	0.63	0.68	0.66	0.76	0.72
	0.01	0.50	0.52	0.52	0.68	0.85	0.74	0.61	0.66	0.64	0.66	0.72	0.72
$K = 10$	0.10	0.82	0.83	0.80	0.84	0.85	0.82	0.72	0.80	0.82	0.77	0.83	0.82
	0.05	0.81	0.79	0.78	0.86	0.83	0.83	0.69	0.75	0.79	0.74	0.81	0.81
	0.025	0.64	0.83	0.74	0.83	0.86	0.78	0.65	0.66	0.71	0.68	0.76	0.76
	0.01	0.50	0.66	0.64	0.78	0.79	0.75	0.66	0.66	0.68	0.68	0.72	0.76
$K = 20$	0.10	0.83	0.84	0.80	0.84	0.86	0.83	0.72	0.80	0.85	0.77	0.82	0.82
	0.05	0.84	0.81	0.79	0.88	0.84	0.84	0.69	0.76	0.79	0.74	0.82	0.82
	0.025	0.71	0.84	0.77	0.84	0.87	0.79	0.64	0.67	0.74	0.69	0.77	0.76
	0.01	0.50	0.64	0.69	0.83	0.81	0.77	0.68	0.68	0.70	0.68	0.73	0.77
Bagged ROSE													
$K = 5$	0.10	0.84	0.86	0.83	0.83	0.85	0.84	0.78	0.81	0.83	0.79	0.79	0.84
	0.05	0.86	0.85	0.85	0.86	0.86	0.85	0.75	0.80	0.81	0.77	0.80	0.80
	0.025	0.80	0.83	0.85	0.82	0.82	0.84	0.67	0.79	0.77	0.68	0.80	0.79
	0.01	0.75	0.82	0.82	0.72	0.82	0.81	0.66	0.76	0.78	0.67	0.74	0.77
$K = 10$	0.10	0.85	0.87	0.85	0.85	0.86	0.85	0.79	0.83	0.87	0.80	0.84	0.86
	0.05	0.87	0.88	0.86	0.86	0.88	0.86	0.77	0.82	0.83	0.78	0.84	0.87
	0.025	0.83	0.85	0.85	0.83	0.85	0.85	0.70	0.80	0.82	0.69	0.79	0.81
	0.01	0.78	0.82	0.82	0.78	0.82	0.82	0.67	0.77	0.80	0.67	0.77	0.80
$K = 20$	0.10	0.85	0.87	0.85	0.85	0.87	0.85	0.80	0.84	0.87	0.80	0.83	0.86
	0.05	0.87	0.87	0.85	0.87	0.87	0.86	0.78	0.83	0.83	0.77	0.83	0.86
	0.025	0.83	0.86	0.85	0.82	0.86	0.85	0.69	0.81	0.83	0.69	0.80	0.82
	0.01	0.78	0.82	0.83	0.78	0.82	0.83	0.69	0.77	0.82	0.65	0.76	0.81

mult is used to define the actual size of the training set on which the estimation of the classifier is based (see Table 3 for details). K is the number of iterations

general, the larger the original training set, the higher the AUC. A tendency towards a depletion in accuracy, when the imbalance increases, is still evident when using both the procedures. Simulations show that, in most the considered scenarios, ROSE outperforms SMOTE and such improvement is mainly evident for extreme levels of imbalance and small sample sizes. In such cases, while the new examples generated by ROSE lie in the elliptical neighborhood of the observed rare data, synthetic examples generated by means of SMOTE lie along the line segments joining the minority class

examples. Thus, the use of SMOTE risks the decision region associated to the rare class in the features space not to be enlarged enough.

The regularization method proposed by Lee (2000) also aids the improvement of the performance of the classifier (Table 6). Here, a larger number of noisy samples generated for a given training set corresponds to higher values of the AUC. Higher levels of accuracy are also associated with a larger number of noisy replicates for each rare example. Again, when the class imbalance gets extreme, the classifier tends to make more mistakes, especially when the training set is small, but such a reduction of accuracy decreases when both the number of noisy samples and the number of noisy replicates for each rare example are large (this is specially evident from the Mixture of Normal data).

Interesting considerations arise from the comparison between Lee's method and ROSE: when both the number of noisy samples and the number of noisy replicates for each rare example are large, ROSE performs, in general, worse than its competitor, and when the imbalance between classes is not extreme, ROSE again cannot do better than Lee (but it should be reminded that a combination of classifiers is used, instead of one, in this instance). However, when the rare examples amount to 1 or 2.5 % of the observed data, the AUCs obtained by applying ROSE are comparable or even larger than the corresponding values obtained when Lee's procedure is applied, even if the computational complexity is much lower.

Being interpretable as a bagging classifier, the iterative application of ROSE (Table 6) outperforms the other considered techniques unsurprisingly. However, it is more efficient than Lee's proposal because a few iterations are enough to offset the effect of a strong imbalance between the classes.

Concerning the real data applications, results reported in Table 7 basically confirm the considerations drawn from the use of simulated data. Absolute imbalance settings (with very few observed minority examples) determine low accuracy of classification if no measure is taken. See, in particular, the *Wine quality*, *Transfusion*, *Glass*, and *Vehicle Silhouette* data. When the number of minority examples is relatively low but still substantial (e.g. the *Forest cover*, *Adult*, *Phoneme* and *Concrete Compressive Strength* data) or the classes are well separated (as in the *Breast Cancer*, *Hypothyroid*, *Vertebral Column* and the *Cardiotocography* data), the effects of the class imbalance are less critical and quite an accurate classification may be obtained. In the former case, direct undersampling improves highly the performance of the classifiers.

Generating new synthetic examples is of greater advantage, with ROSE outperforming the other remedies in most of situations. Moreover, the gain of using ROSE, instead of other remedies, tends to increase with the class imbalance (see, in particular, the results of the four most imbalanced data sets). SMOTE, on the other hand, also results very effective. However, the AUCs obtained by using ROSE are greater than the SMOTE's ones in 18 of the 20 considered scenarios. The method of Lee (2000), instead, performs on real data worse than on simulated data, even deteriorating the accuracy of the classifier estimated without remedies to the imbalance on several data sets.

In order to evaluate if the accuracy of the considered remedies to imbalance is significantly different, a Friedman test has been performed. The greatest accuracy of ROSE has been then confirmed by a pairwise comparison between ROSE and the other

Table 7 AUC obtained on real data with classification trees and different remedies to the class imbalance

Data	Imbalanced data	Undersampling	SMOTE	Lee regularization	ROSE
Wine quality	0.557	0.604	0.624	0.600	0.782
Forest cover	0.785	0.855	0.889	0.878	0.903
Infocamere	0.536	0.567	0.781	0.514	0.768
Abalone	0.571	0.714	0.663	0.601	0.717
Hypothyroid	0.658	0.927	0.959	0.823	0.974
Adult	0.624	0.796	0.752	0.589	0.794
Phoneme	0.758	0.726	0.751	0.781	0.828
Breast cancer	0.822	0.886	0.879	0.919	0.898
Cardiotocography	0.673	0.795	0.848	0.694	0.852
Transfusion	0.558	0.584	0.561	0.598	0.660
Glass	0.533	0.533	0.712	0.512	0.760
Pima indians	0.571	0.583	0.665	0.575	0.714
Cylinder bands	0.571	0.539	0.534	0.524	0.644
Vehicle silhouettes	0.551	0.536	0.521	0.571	0.698
Image segmentation	0.789	0.877	0.906	0.511	0.904
Spectf	0.574	0.668	0.565	0.683	0.685
Vertebral column	0.686	0.816	0.816	0.914	0.851
Parkinsons	0.629	0.725	0.719	0.789	0.752
Concrete compressive strength	0.799	0.826	0.804	0.615	0.819
Credit screening	0.682	0.757	0.741	0.792	0.759
p values	$<10^{-3}$	0.001	$<10^{-3}$	0.002	–

The last line lists the p values associated to testing the pairwise difference between ROSE and the other methods, over the considered data sets

methods, after adjusting the obtained p values to control for the familywise error rate. The tests adopted are described by Demsar (2006). The use of logit models (instead of classification trees) generally leads to higher levels of accuracy, but analogous considerations about the comparative behaviour of ROSE and the other methods may be drawn.

4.2 Model evaluation by using ROSE

In Section 2.2, it was outlined that creating new artificial examples from a conditional density estimate of the two classes gives the opportunity to exploit the observed data to test the accuracy of the estimated classification rule. Now, we give an illustration about the soundness of this practice.

Simulations have been conducted by generating data from the two structures mentioned in Table 3. Compared to the simulation performed in the previous section, the cardinalities of the two classes are not fixed, but depend on a predetermined probability. This choice is due to the necessity of taking into account the variability of the data (and, hence, also the sizes of the classes) to obtain reliable estimates of the classifier

accuracy. In particular, the probabilities of the the rare class have been chosen in the set {0.5, 0.1, 0.01}.

Moreover, the opportunity of using ROSE in assessing the model has been also tested on two real data sets, i.e. the Adult data and the Forest Cover data. The choice of these two data sets only among the 20 considered ones is motivated by the need of using very large sample sizes to compute the "true" accuracy of the classifier and testing the accuracy's estimators.

Again, classification trees and logit models are the used learning methods.

The area under the ROC curve (AUC) has been chosen as an evaluation metric for the analysis. Three methods for estimating the AUC have been assessed: the resubstitution method, consisting of measuring the accuracy of the classifier on the training set, the holdout method, where the available data are split into a training set and a test set, and the practice of using the observed data for testing the classifier after that artificial data generated by ROSE have been used for the training stage.

The simulation design follows several previous works aimed at evaluating the performance of different error estimators (Chernick et al. 1985; Wehberg and Schumacher 2004). The number of simulation trials have been set to 100. For each simulation trial, a sample of size 1,000 is drawn and used as follows: for the resubstitution method, the sample is directly employed to both estimate the classification rule and test it; concerning the holdout method, a random 75 % of the sample is used to train the classifier and the remaining 25 % is used to test it; finally, a ROSE artificial training set is generated from the selected sample, which, in turn, serves as a test set. In each case the "true" AUC (conditional on the training set) is approximated by testing the classifier on 1,000 samples of size 10,000 drawn from the same population as the training samples and averaging the resulting AUCs. The true AUCs associated to the real data examples have been built by using 10,000 data not involved in the model estimation. The true AUCs are computed for each simulation trial. The bias of the three estimators of the AUC is obtained by averaging the differences between the true AUC and the corresponding estimates computed for each of the simulation trials. Moreover, the standard deviation of the estimates has been computed, and the root mean square of the differences between the true AUC and the estimates has been used as a summarizing measure of estimator performance. Results are reported in Tables 8 and 9.

Not surprisingly, the apparent AUC provides an optimistic estimate of the true AUC, if the prediction procedure is highly data-dependent (classification tree). Moreover, it is clear that the more imbalanced the distribution of the classes is, the more biased the estimate of the AUC is, when the resubstitution method is used. If a less data-dependent procedure is used for prediction, e.g. the logit model, the tendency of the resubstitution method to overestimate the true AUC is less remarkable.

The holdout method would be supposed to provide better estimates of the classifier's accuracy. In fact, while it appears reasonably unbiased, it suffers from high variability as the skewness in the distribution of the classes increases. This behavior occurs regardless of the considered classifier, thus making this estimator totally inadequate for use in a context of imbalanced learning.

The practice of testing the accuracy on the originally observed data, after training the classifier on synthetic examples generated according to ROSE, appears to be

Table 8 Simulated data: bias, standard error and root mean squared error of three methods for estimating the AUC: the resubstitution method, the holdout method, and the practice of using the observed data for testing the classifier after that artificial data generated by ROSE have been used for the training stage

Mixture of normal data	50 %	10 %	1 %
Classification tree			
BIAS			
Resubstitution	0.031	0.114	0.412
Holdout	0.002	0.001	−0.012
ROSE	0.007	0.010	0.058
SD			
Resubstitution	0.018	0.063	0.019
Holdout	0.029	0.095	0.114
ROSE	0.025	0.030	0.057
RMSE			
Resubstitution	0.033	0.123	0.419
Holdout	0.027	0.068	0.111
ROSE	0.016	0.023	0.088
Logit model			
BIAS			
Resubstitution	0.000	0.002	0.032
Holdout	−0.003	0.000	0.008
ROSE	0.000	0.002	0.025
SD			
Resubstitution	0.011	0.014	0.046
Holdout	0.020	0.030	0.136
ROSE	0.011	0.014	0.040
RMSE			
Resubstitution	0.011	0.015	0.073
Holdout	0.020	0.030	0.132
ROSE	0.011	0.014	0.052
Filled semi-hypersphere			
Classification tree			
BIAS			
Resubstitution	0.027	0.109	0.201
Holdout	−0.004	−0.009	−0.013
ROSE	0.011	0.027	0.079
SD			
Resubstitution	0.062	0.021	0.051
Holdout	0.036	0.055	0.144
ROSE	0.040	0.022	0.034
RMSE			
Resubstitution	0.036	0.119	0.220
Holdout	0.024	0.039	0.114
ROSE	0.018	0.031	0.087

Table 8 continued

Filled semi—hypersphere	50 %	10 %	1 %
Logit model			
BIAS			
Resubstitution	0.004	0.003	0.016
Holdout	0.001	−0.009	0.004
ROSE	0.005	0.008	0.011
SD			
Resubstitution	0.009	0.018	0.022
Holdout	0.011	0.054	0.183
ROSE	0.009	0.016	0.041
RMSE			
Resubstitution	0.012	0.017	0.037
Holdout	0.012	0.056	0.172
ROSE	0.013	0.019	0.036

unquestionably winning among the considered estimators of the AUC. Indeed, the bias of the estimates generally exceeds the bias of the holdout method, and both the bias and the variance of the estimates show a clear tendency to increase as the class imbalance gets more extreme, but the root mean square error of the estimates results the lowest one, whatever the level of skewness in the class distribution is and whatever the considered classifier is used.

However, two main arguments have to be remarked, prompting that any conclusion about the conducted simulation should be drawn cautiously. First, the true AUC is not a constant but a random variable which varies within different training sets. Hence, the relation between bias, variance and the root mean square error does not hold in this context (Chernick et al. 1985). Secondly, a reliable interpretation of results would require that different sources of variation of the results were kept separated.

Nonetheless, the three mentioned methods for estimating the AUC cannot be evaluated on equal terms: when the resubstitution method and ROSE are considered, the observed 1000−sized sample is used as a test set and 1, 000 examples are involved in training the classifier. In contrast, only 250 observations serve to test the model when the holdout method is used, and the remaining 750 data are employed for the training stage. Disparity of such conditions could be a reason for explaining, for instance, why ROSE seems to outperform the holdout method even when the classes are balanced.

Moreover, we cannot know if the quality of the accuracy estimate is independent of the quality of the classifier: it is not to exclude that better estimates of the AUC are associated with more predictive learners. However, given that the training stage and the evaluation of the classifier are inseparable, we have adhered to the conditions occurring when one faces a real data problem of classification. In such contexts, given the available data, the best method is the one that strikes the balance between quality of prediction and goodness of the estimate of such quality.

These arguments along with the results in Sect. 4.1 suggest that an uncritical application of the discussed validation method cannot be blindly recommended. On the

Table 9 Real data: bias, standard error and root mean squared error of the resubstitution method, the holdout method, and the practice of using the observed data for testing the classifier after that artificial data generated by ROSE have trained it

Adult data	50 %	10 %	1 %
Classification tree			
BIAS			
Resubstitution	0.020	0.089	0.401
Holdout	0.003	−0.006	0.012
ROSE	0.006	0.011	0.083
SD			
Resubstitution	0.012	0.028	0.019
Holdout	0.019	0.068	0.138
ROSE	0.021	0.019	0.038
RMSE			
Resubstitution	0.021	0.093	0.407
Holdout	0.015	0.042	0.122
ROSE	0.010	0.016	0.087
Logit model			
BIAS			
Resubstitution	0.005	0.013	0.139
Holdout	0.004	0.004	0.080
ROSE	0.004	0.007	0.079
SD			
Resubstitution	0.007	0.015	0.058
Holdout	0.016	0.032	0.145
ROSE	0.007	0.012	0.030
RMSE			
Resubstitution	0.009	0.021	0.153
Holdout	0.017	0.031	0.169
ROSE	0.009	0.015	0.095
Forest cover data			
Classification tree			
BIAS			
Resubstitution	0.017	0.031	0.285
Holdout	0.004	0.005	−0.029
ROSE	0.007	0.016	0.094
SD			
Resubstitution	0.020	0.064	0.085
Holdout	0.026	0.070	0.146
ROSE	0.054	0.048	0.081
RMSE			
Resubstitution	0.020	0.039	0.299
Holdout	0.020	0.041	0.142
ROSE	0.014	0.025	0.128

Table 9 continued

Forest cover data	50 %	10 %	1 %
Logit model			
BIAS			
Resubstitution	0.002	0.004	0.075
Holdout	−0.001	0.002	0.068
ROSE	0.002	0.006	0.070
SD			
Resubstitution	0.011	0.020	0.054
Holdout	0.021	0.045	0.148
ROSE	0.011	0.018	0.047
RMSE			
Resubstitution	0.011	0.022	0.106
Holdout	0.021	0.045	0.172
ROSE	0.011	0.019	0.099

one hand, the variance of the holdout method might explode, especially when the imbalance between classes is both extreme and absolute (with a very few observed positive examples). If a repeated random split of the data into training and test sets gives this indication, the joint use of ROSE for training and assessing the classifier has been proven to be helpful. On the other hand, it should be borne in mind that this practice risks to lead to an over-optimistic evaluation of the accuracy, due to the bias increase and, *ceteris paribus*, a holdout method should be considered as the most reliable choice to understand the classifier ability of generalization.

4.3 ROSE in practice

In this section, we show an illustration of how to use ROSE operationally for both estimation and evaluation of a classifier. In the light of the arguments previously remarked, we consider here only the data sets presenting characteristics of both extreme and absolute imbalance between the classes: the *Wine quality*, *Abalone*, *Hypothyroid*, *Breast Cancer*, *Transfusion*, *Glass*, *Pima Indians*, *Cylinder Bands*, *Vehicle Silhouettes*, *Spectf*, *Vertebral Column*, *Parkinson* data sets.

In exploiting ROSE for evaluation purposes, different practices can be chosen: an holdout version of ROSE, for instance, would consist in testing the classifier on the original data after training it on a balanced ROSE sample. Alternatively, bootstrap or cross-validated versions of ROSE may be considered, as illustrated in Table 10. Here, we show an application of the use of the bootstrap version of ROSE.

The classifiers have been trained on 50 balanced ROSE samples generated from each data set and the performance of the estimated classification rules have been evaluated by measuring the AUC on the originally observed data. As a benchmark, the estimation of the classifiers on 50 imbalanced smoothed bootstrap samples drawn from the same data sets has been considered. The obtained empirical distributions of the AUC when the training samples vary have been reported in Figs. 1 and 2.

Table 10 Alternative ways to use ROSE for model assessment

Leave-one-out ROSE	Bootstrap ROSE
for (i: = 1 to 10) do	**for** (b: = 1 **to** B) do
get a ROSE balanced sample $T^*_{m(i)}$ from$T_n \setminus (\mathbf{x}_i, y_i)$	
	get a ROSE balanced sample $T^*_{m(b)}$ from T_n
estimate $R_{T^*_{m(i)}}$	estimate, $R_{T^*_{m(b)}}$
test the accuracy of $R_{T^*_{m(i)}}$ on $(\mathbf{x}_i, y_i)$	test the accuracy of $R_{T^*_{m(b)}}$ on T_n
end for	**end for**
	get the bootstrap distribution of the accuracy measure.

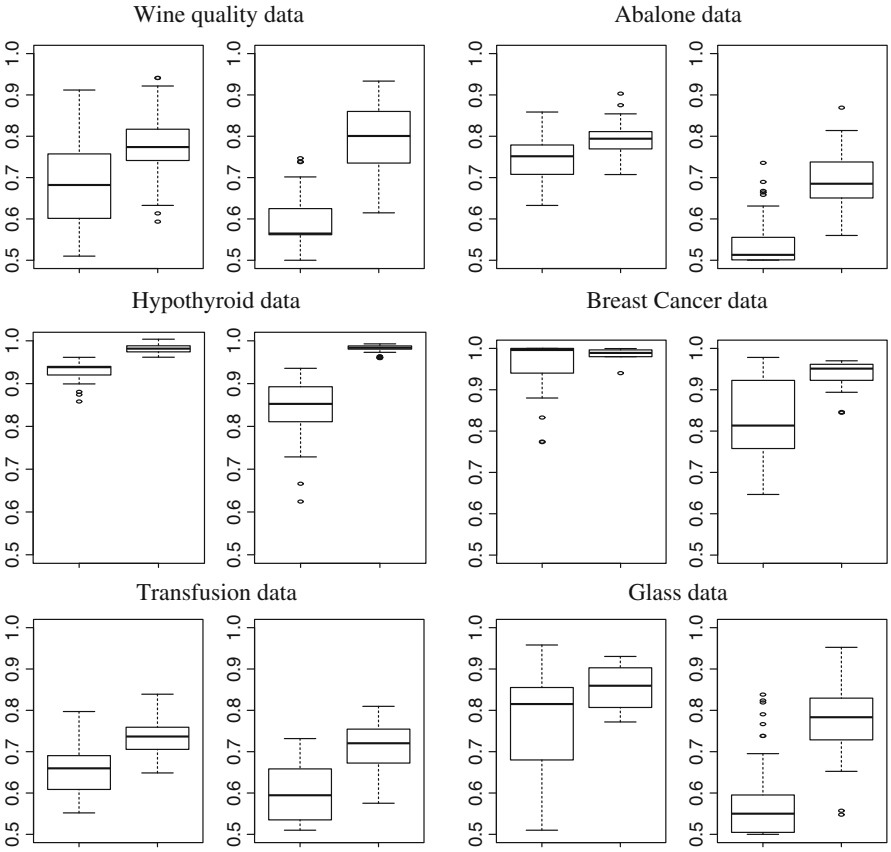

Fig. 1 For each data set, the AUC distribution is plotted when logit models (*left*) and classification trees (*right*) are used. In each panel, the first boxplot displays the AUCs obtained by training the classifier on the smoothed bootstrap imbalanced samples, and the the second boxplot results from training the classifier on ROSE samples

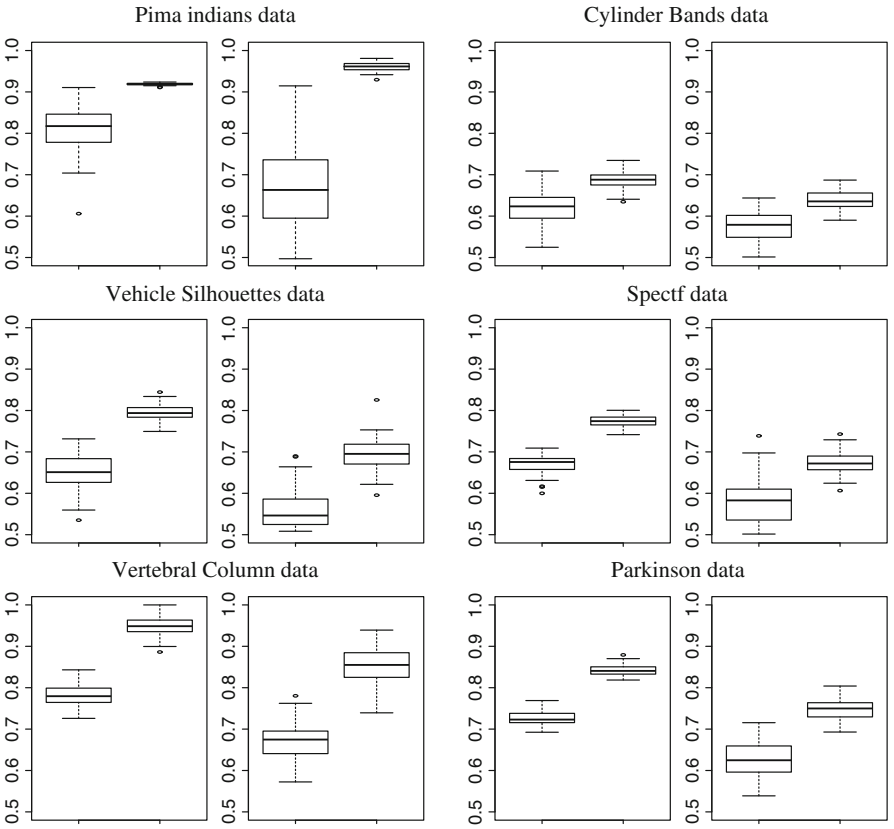

Fig. 2 Distribution of the AUC. Cf. Fig. 1 for further details

When classification trees are used to learn from data and no remedy is adopted for coping with the class imbalance, there is a high risk of producing rules not much more accurate than random guess. Indeed, the median AUC is not that low, but the variability of the AUC distributions is high and the inferior whiskers of the plots brush against the value of 0.5 in most the applications. Exceptions are the *Breast Cancer*, *Hypothyroid*, *Parkinson* and the *Vertebral Column* data, where the classification task is apparently simple despite the imbalance. This is also consistent with resuts in Sect. 4.1. When ROSE is run prior to the tree building, the dispersion of the AUC is often lower than the corresponding dispersion if the class imbalance is ignored, and the median AUC always exceeds the median AUC resulting from training the tree on imbalanced data. Excellent results are obtained when the *Hypothyroid* and the *Pima Indians* data sets are examined, since ROSE manages to get an almost perfect prediction. Interestingly, ROSE can improve the accuracy of the learner even when the classification task is simple and satisfactory results are obtained without resorting to any remedy.

Ignoring the class imbalance is less risky when a logit model is used. The range of the AUC distributions shifts towards remarkably higher values than the distributions associated with the use of classification trees. Moreover, the variability of such

distributions is perceptibly lower. However, the gain in applying ROSE before model estimation is even larger, and prediction of classifiers trained on ROSE samples almost uniformly outperforms predictions based on imbalanced data.

It is interesting to note that, although decision trees are the most frequently used classifiers in imbalanced learning, the examples reported throughout the paper clearly show an undisguised superiority of the logit model, whose performance appears either more accurate and more precise.

5 Final remarks

In this work, we have provided a comprehensive study about the use of imbalanced data in binary classification: the main causes of the failure of both parametric and non-parametric standard classifiers have been reviewed, and some new perspectives have been discussed about the effects of class imbalance. Inherent literature has grown at an explosive rate in recent years, but it has mainly focused on proposing sophisticated learning methods or alternative evaluation metrics. Instead, the problem of high variability of the accuracy estimator has been totally ignored. In fact, when the distribution of the classes is skewed, the estimated models perform very poorly and bad estimates of the classifier performance may lead to misleading conclusions on the quality of the prediction.Stemming from the need of dealing simultaneously with the problems of model estimation and model evaluation, we have discussed a unified and systematic framework, hinging on a smoothed bootstrap form of data re-sampling.

The proposed technique includes the existing solutions based on oversampling as a special case; it is supported by a theoretical framework and reduces the risk of model overfitting. The application of the proposed technique to real and simulated data has shown excellent performance, compared with other similar methods already known in the literature. Further refinements are still possible and in the light of the promising results diffusely obtained in the direction of data augmentation, looking for alternative methods to generate synthetic data will be one of the focus of future research. As a guiding line, while effective criteria should firstly follow the data driven paradigm, it seems advisable that new advances should rely upon use of theoretically sound arguments.

A further advantage of the proposed methodology is the opportunity of combining creation of synthetic data with ensemble ideas. We have shown that ROSE is strictly related to bootstrap methods allowing its natural extension in terms of bagging learner, which claims improved performance of classification. Moreover, smoothed boosting can be easily carried on in conjunction with ROSE by relating the weights update mechanism to the probability of data generation. This is a project by itself and will be object of further investigation.

The proposed technique may also aid improving the estimate of the learner accuracy, especially in those situations of extreme absolute imbalance, when the reliability of standard estimators of accuracy decreases. In spite of the promising results, the need of some caution in evaluating the model accuracy should be always bore in mind in the context of imbalanced learning. Moreover, the effects of additional sources of complication such as class overlapping and data fracture may play a critical role on

the bias and variability of the accuracy estimator. These issues, along with the study of the behavior of more refined estimators of accuracy, need a deeper understanding and will be object of future research.

Acknowledgments The authors wish to thank the anonymous reviewers for their fruitful comments that greatly improved the presentation of this paper.

References

Akbani R, Kwek S, Japkowicz N (2004) Applying support vector machines to unbalanced datasets. In: Boulicaut JF, Esposito F, Giannotti F, Pedreschi D, eds. Lecture Notes in Computer Science, Proceedings of 15th European conference on machine learning, ECML, Springer, Pisa, 3201:39–50

Asuncion A, Newman DJ (2007) UCI machine learning repository http://www.ics.uci.edu/~mlearn/MLRepository.html. University of California, School of Inf. and Comput. Sci., Irvine

Barandela R, SÃnchez JS, GarcÃá1a V, Rangel E (2003) Strategies for learning in class imbalance problems. Patt Recognit 36:849–851

Batista G, Prati R, Monard M (2004) A study of the behaviour of several methods for balancing machine learning training data. SIGKDD Explor 6(1):20–29

Batuwita R, Palade V (2010) FSVM-CIL: fuzzy support vector machines for class imbalance learning. IEEE Trans Fuzzy Syst 18(3):558–571

Bowman AW, Azzalini A (1997) Applied smoothing techniques for data analysis: Kernel approach with S-plus illustrations. Oxford University Press, Oxford

Breiman L (1996) Bagging predictors. Mach Learn 24:123–140

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth International Group, Belmont, CA

Burez J, Vanden Poel D (2009) Handling class imbalance in customer churn prediction. Expert Syst Appl 36:4626–4636

Chawla NV (2003) 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 Class Imbalances

Chawla NV, Bowyer KW, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357

Chernick M, Murthy V, Nealy C (1985) Application of bootstrap and other resampling methods: evaluation of classifier performance. Pattern Recogn Lett 3:167–178

Cieslak D, Chawla N (2008) Learning decision trees for unbalanced data. Lect. Notes in Comput. Sci., 5211:241–256

Cramer JS (1999) Predictive performance of binary logit models in unbalanced samples. The Statistician 48:85–94

Davis J, Goadrich M (2006) The relationship between Precision-Recall and ROC curves. In: Cohen W, Moore A, eds. Proceedings of the 23rd International Conference on Machine Learning, ACM Press, Pittsburgh, PA, pp 233–240

Demsar J (2006) Statistical comparison of classifiers over multiple data sets. J Mach Learn Res 7(7):1–30

Drummond C, Holte RC (2006) Cost curves: an improved method for visualizing classifier performance. Mach Learn 65(1):95–130

Efron B, Tibshirani R (1993) An introduction to the bootstrap. Chapman and Hall, New York

Eitrich T, Kless A, Druska C, Meyer W, Grotendorst J (2007) Classification of highly unbalanced CYP450 data of drugs using cost sensitive mach learning techniques. J Chem Inform Model 47(1):92–103

Estabrooks A, Taeho J, Japkovicz N (2004) A multiple resampling method for learning form imbalanced data sets. Comput Intell 20:18–36

Fernandez A, Barrenechea E, Bustince H, Herrera F (2012) A review on ensembles for the class imbalance problem: bagging, boosting, and hybrid-based approaches. IEEE Trans Syst, Man, Cybern, C 42:463–484

García S, Derrac J, Triguero I, Carmona CJ, Herrera F (2012) Evolutionary-based selection of generalized instances for imbalanced classification. Knowl Based Syst 25:3–12

Guo H, Viktor HL (2004) Boosting with data generation: improving the classification of hard to learn examples. SIGKDD Explor 6(1):30–39

Hand D (2006) Classifier technology and the illusion of progress. Stat Sci 21(1):1–14
Hand D, Vinciotti V (2003) Choosing K for two-class nearest neighbour classifiers with unbalanced classes. Patt Recognit Lett 24:1555–1562
He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng, 21(9)
Japkowicz N, Stephen S (2002) The class imbalance problem: a systematic study. Intell Data An J 6
Jo T, Japkowicz N (2004) Class imbalances versus small disjuncts. SIGKDD Explor 6(1):40–49
Khoshgoftaar TM, Golawala M, Van Hulse J (2007) An empirical study of learning from imbalanced data using random forest. Proceedings of the 19th IEEE international conference on tools with artif intelligence, vol 2, Washington, DC
Khoshgoftaar TM, Van Hulse J, Napolitano A (2011) Comparing boosting and bagging techniques with noisy and imbalanced data. IEEE Trans on Syst, Man, Cybern.-Part A: Syst Humans 41(3):552–568
King EN, Ryan TP (2002) A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression. Am Stat 56:163–170
King G, Zeng L (2001) Logistic regression in rare events data. Political Anal 9:137–163
Kotsiantis S, Kanellopoulos D, Pintelas P (2006) Handling imbalanced datasets:a review. GESTS International Transactions on Computer Science and Engineering, vol 30
Kukar M, Kononenko I (1998) Cost-sensitive learning with neural networks. Proceedings of the 13th European conference on artificial intelligence, Wiley, New York, pp 445–449
Kubat M, Matwin S (1997) Addressing the curse of imbalanced training sets: one-sided selection. Proceedings of the 14th international conference on machine learning. ICML, Nashville, pp 179–186
Lee S (2000) Noisy replication in skewed binary classification. Comput Stat Data An 34:165–191
Lee S (1999) Regularization in skewed binary classification. Comput Stat 14:277–292
Lin Y, Lee Y, Wahba G (2002) Support vector machines for classification in nonstandard situations. Mach Learn 46:191–202
Liu Y, Chawla NV, Harper MP, Shriberg E, Stolcke A (2006) A study in machine learning from imbalanced data for sentence boundary detection in speech. Comput Speech & Lang 20:468–494
Mazurowski MA (2008) Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural Netw 21:427–436
McCarthy K, Zabar B, Weiss G (2005) Does cost-sensitive learning beat sampling for classifying rare classes? Proceedings of the 1st international workshop on utility-based data mining, ACM Press, New York, pp 69–77
Mease D, Wyner A, Buja A (2007) Boosted classification trees and class probability-quantile estimation. J Mach Learn Res 8:409–439
Oommen T, BaiseL Vogel R (2011) Sampling bias and class imbalance in maximum-likelihood logistic regression. Math Geosci 43:99–120
Pavón R, Laza R, Reboiro-Jato M, Fdez-Riverola F (2011) Assessing the impact of class-imbalanced data for classifying relevant/irrelevant medline documents. Adv Intell Soft Comput 93:345–353
Percannella G, Soda P, Vento M (2011) Mitotic HEp-2 cells recognition under class skew. Lecture Notes in Computer Science (including Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp 353–362
Riddle P, Segal R, Etzioni O (1994) Representation design and brute-force induction in a Boeing manufacturing domain. Appl Artif Intell 8:125–147
Schiavo RA, Hand DJ (2000) Ten more years of error rate research. Int Stat Rev 68(3):295–310
Silverman BW (1986) Density estimation for statistics and data analysis. Chapman and Hall, New York
Ström F, Koker R (2011) A parallel neural network approach to prediction of Parkinson's Disease. Expert Syst Appl 38(10):12470–12474
Sun Y, Kamel MS, Wong AKC, Wang Y (2007) Cost-sensitive boosting for classification of imbalanced data. Patt Recogn 40(12):3358–3378
Sun Y, Wong AKC, Kamel MS (2009) Classification of imbalanced data: a review. Int J Patt Recogn Artif Intell 23(4):687–719
Ting KM (2002) An instance-weighting method to induce cost-sensitive trees. IEEE Trans Knowl Data Eng 14(3):659–665
Thomas J, Jouve P, Nicoloyannis N (2006) Optimisation and evaluation of random forests for imbalanced datasets. Lecture Notes in Computer Science, Springer 4203:622–631
Veropoulos K, Campbell C, Cristianini N (1999) Controlling the sensitivity of support vector machines. Proceedings of the international joint conference on artificial intelligence, Stockholm, pp 55–60

Wasikowski M, Chen XW (2010) Combating the small sample class imbalance problem using feature selection. IEEE Trans Knowl Data Eng 22(10):1388–1400

Wehberg S, Schumacher M (2004) A comparison of nonparametric error rate estimation methods in classification problems. Biom J 46(1):35–47

Weiss GM (2004) Mining with rarity: a unifying framework. ACM SIGKDD Explor. Newsletter 6(1)

Weiss GM, Provost F (2001) The effect of class distribution on classifier learning: an empirical study. Technical report, ML-TR-44, Department of Computer Science, Rutgers University, New Jersey

Wu XLJ, Zhou Z (2009) Exploratory undersampling for class-imbalance learning. IEEE Trans: On Syst., Man, Cybern., B 39:539–550

Yen S, Lee Y (2006) Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. Intelligent Control and Automation. Series: Lecture Notes in Control and Information Sciences, pp 731–740

Zhou Z, Liu X (2006) Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Trans Knowl Data Eng 18(1):63–77