Dipartimento di scienze economiche,
aziendali, matematiche e statistiche
“Bruno de Finetti”

Working Paper Series, N. 2, 2010

Training and assessing classification rules with unbalanced data

GIOVANNA MENARDI
Department of Statistical Sciences
University of Padova

NICOLA TORELLI
Department of Economics, Business, Mathematics and Statistics "Bruno de Finetti"
University of Trieste

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Working paper series

Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche
"Bruno de Finetti"
Piazzale Europa 1
34127, Trieste
Tel. ++40 558 7927
Fax ++40 567543
http://www.deams.units.it

EUT Edizioni Università di Trieste
Via E.Weiss, 21 - 34128 Trieste
Tel. ++40 558 6183
Fax ++40 558 6185
http://eut.units.it
eut@units.it

ISBN: 978-88-8303-321-6

Working Paper Series, N. 2, 2010

Training and assessing classification rules with unbalanced data

GIOVANNA MENARDI
Department of Statistical Sciences
University of Padova

NICOLA TORELLI
Department of Economics, Business, Mathematics and Statistics "Bruno de Finetti"
University of Trieste

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 both the problems is proposed, based on a smoothed bootstrap re-sampling technique.

KEYWORDS: accuracy, binary classification, bootstrap, kernel density estimation, unbalanced learning.

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 couple ($\mathbf{x}$, 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 $T_n = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_i, y_i), \ldots, (\mathbf{x}_n, y_n)$ of such couples (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 the ratio between the estimated conditional probability of belonging to $\mathcal{Y}_j$ and the estimated conditional probability of belonging to another group exceeds a given threshold t, typically set to 1:

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

Several techniques have been proposed in the literature for dealing with the classification task: from the more heuristic approaches typical of classification trees or nearest neighbors, to the traditional discriminant analysis and multinomial models and the more complex support vector machines, neural networks or ensemble techniques. These classification methods are basically characterized by some implicit or explicit approach to the estimation of the unknown probabilities involved in Equation 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 behaviours) and computer sciences (feature recognition in image data).

In certain domains (like those just mentioned), the class imbalance is intrinsic to the problem. However, unbalanced 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 (Japkowicz and Stephen, 2002).

A massive interest in unbalanced 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 problem. 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, 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, and that aspects of real practical problems often render such small differences irrelevant, or even unreal, so that the gains reported on theoretical grounds [...] do not translate into real advantages in practice". In unbalanced learning, 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.

Moreover, while the literature has fully addressed the issues of model estimation and choice of the accuracy metric in unbalanced learning, a critical inherent aspect has been completely ignored by the research community. Whatever metric is chosen for measuring the classifier's 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 has not been considered yet. The purpose of this work is to address such an issue, by providing a unified and systematic framework for dealing with unbalanced learning both from the perspective of model training and model evaluation. The proposed technique, referred to as ROSE (Random OverSampling Examples), is based on a smoothed bootstrap form of re-sampling from data.

Section 2 discusses the effects of a highly skewed distribution of the classes when building and measuring the accuracy of classification rules. In Section 3, our contribution is presented, the properties of the proposed technique are enlightened and a comparison to some similar existing remedies is conducted. Section 4 presents results from simulations and from the application of the proposed technique to some real data sets, aimed at showing that ROSE may be effectively used to estimate the accuracy of a classifier . Some final remarks conclude the paper.

2. The effects of class imbalance

In many practical applications of binary classification problems, an extremely unbalanced 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), 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. However, the current section aims at understanding the main issues that emerge in modeling and assessing the accuracy of unbalanced data.

a. 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 such a 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 unbalanced, because the conditional probabilities of the rare class are underestimated (King and Zeng, 2001). The use of logistic regression in classification problems with skewed data is also discussed in Cramer (1999), Owen (2007) and King and Ryan (2002).

The performance of linear discriminant analysis is also compromised when the distribution of the classes is unbalanced. 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) and has further given rise to a heated debate in Xie and Qiu (2007) and Xue and Titterington (2008).

Not even the more flexible nonparametric methods are immune to the consequences of a skewed distribution of the classes. Basically, such 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 favour classification rules that perform well only on the frequent class (see the next section for further details about using overall accuracy measures in unbalanced learning). Classification trees, for instance, are grown by finding successive divisions such that the decrease in impurity is maximized. This is typically translated either in trivial models having a high accuracy on the prevalent class and a very low accuracy in classifying the rare events or in complex trees that typically overfit the training data (for a discussion about the use of classification trees in an unbalanced framework see, for instance, Chawla, 2003; Cieslak and Chawla, 2008).

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 "as the number of negative examples in a noisy domain grows (the number of positives being constant), so does the likelihood that the nearest neighbor of any example will be negative". However, 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. In such a case, they suggest that the best classification rule for predicting the rare class is based on $k = 1$. 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.

Other commonly used nonparametric classifiers show performance compromised by the

unbalance of the classes. Lawrence et al. (1998) show that neural networks are ineffective because the theoretical assumptions do not hold in the presence of rare events, while Akbani et al. (1999) explain the failure of support vector machines when there is a high degree of imbalance between the classes. Basically, support vector machines minimize the empirical error while maximizing the margin (which is related to the complexity of the classification rule and the ability of generalization), but the imbalance between classes sets this trade-off apart by allocating all the rare examples to the prevalent class (thus making the empirical error very low) and enlarging the margin.

A quite recent argument that moves away from the ones just mentioned suggests that the consequences of class imbalance are not directly caused by the distribution of the classes, but rather that class imbalance may leads to some small disjuncts that, in turn, determine degradation of the classification (Jo and Japkowicz, 2004).

Most of the current research on classification with unbalanced classes focuses on proposing solutions for improving the stage of model estimation. The literature is wide and the provided remedies are various (see, for a review, Kotsiantis et al., 2006; He and Garcia, 2009). However, the main contributions can be basically summarized in solutions at the learning level and solutions at the data level.

i. Solutions at the learning level aim at strengthening the learning process towards the minority class. A first approach to this class of methods produces some modification of the classifier in order to compensate the imbalance. This approach is generally applied to classifiers whose training is based on the optimization of some 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.

   Riddle et al. (1994), for instance, learn from the positive observations only when building decision trees. A similar, but not so extreme, approach is followed by Cieslak and Chawla (2008) who consider the use of the Hellinger distance as an alternative splitting criterion, less sensitive to the skewed distribution of the classes.

   Several strategies have been proposed for biasing the learning process in support vector machines, e.g. by pushing the hyperplane further away from the rare examples. In Wu and Chang (2005), the kernel matrix is adjusted for better fitting the training data, while different penalty constants are used in Veropoulos et al. (1999) for the two classes.

   Barandela et al. (2003) propose a weighted distance function, to be used in $k-$nearest neighbor classifiers, that gives more weight to the majority class. In this way, the distance to the rare examples is reduced more than the distance to the prevalent examples so that the likelihood that the nearest neighbors of any example are positive increases.

   Other remedies addressing the inadequacy of standard classifiers in the presence of rare events by 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. In general, this approach is followed when the skew distribution of the classes is associated with an unbalanced distribution of the misclassification costs (typically, the cost of misclassifying a rare example is higher than the corresponding cost for an

example belonging to the prevalent class). 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) or cost-sensitive pruning schemes may be applied (see, for example, Draper et al., 1994; Ting, 2002; Bradford et al., 1998). Cost-sensitive neural network have also been widely studied in the context of unbalanced classes. Information about the different costs of misclassification is introduced without altering the process of learning by modifying the output of the network as well as the estimated conditional probabilities of belonging to the classes so that the expected costs of misclassification decrease. Alternatively, instead of minimizing the squared error, the back propagation learning procedure can minimize the misclassification costs. For further discussion, see, for example, Kukar and Kononenko (1998). More examples of modifications of existing learning methods for dealing with different misclassification costs can be found in 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 a different propensities toward misclassification errors consists of moving the classification threshold in Equation 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 algorithmic levels also include the use of combinations of classifiers, by following logics typical of boosting, bagging or random forest. Some references are Sun et al. (2007); Fan et al. (1999); Liu et al. (2006); Thomas et al. (2006); Khoshgoftaar et al. (2007).

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.

ii. Solutions at the data level for dealing with unbalanced classes basically focus on altering the class distribution in order to get a more balanced sample.

Remedies following this approach include various techniques to sample the data. The most common techniques are 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). 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; Barandela et al., 2003; Yen and Lee, 2006).

Indeed, both oversampling and undersampling decrease the overall level of class imbalance, thereby improving the overall accuracy of the classifier. The reason that altering the class distribution of the training data aids learning with highly skewed datasets is that it effectively imposes non-uniform misclassification costs. This equivalence between altering the class distribution of the training data and moving the misclassification cost ratio 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 novel techniques of data generation in Huo 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.

b. 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

TABLE 1. Confusion matrix of a binary classification problem

		predicted	
		0	1
actual	0	TN	FP
	1	FN	TP

for labeling new unobserved examples, depend on our ability to measure the classification accuracy.

Although the literature about model assessing 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. Basically, two problems arise in model assessment in the presence of unbalanced classes concerning the choice of the evaluation measure and the estimate of such a measure of accuracy.

i. 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 (see Table 1). In order to provide comprehensive assessments of unbalanced 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, separately evaluated in the two classes.

   One of the most frequently used tools for evaluating the accuracy of a classifier in the presence of unbalanced 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 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).

ii. 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 unbalanced 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 Equation 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 respectively. 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 unbalanced 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.

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 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 based on altering the distribution of the classes in order to get a balanced sample 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 Random Over Sampling Examples(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, that is $P(\mathbf{x}) = f(\mathbf{x})$ is a probability density function on $\mathcal{X}$. Without loss of generality, we may consider that $n_j < n$ is the size of $\mathcal{Y}_j, j = 0, 1$. The ROSE procedure for generating one new artificial example consists of the following steps:

i. select $y = \mathcal{Y}_j \in \mathcal{Y}$ with probability $\frac{1}{2}$

ii. select $(\mathbf{x}_i, y_i)$ in $\mathbf{T}_n$ such that $y_i = y$ with probability $p_i = \frac{1}{n_j}$

iii. 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 $\mathbf{H}_j$ a matrix 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 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 $f(\mathbf{x}|\mathcal{Y}_j)$.

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, while the originally observed data remain free of being used for testing the classifier. Alternatively, cross-validation or smoothed bootstrap methods could be used. It is worthwhile to note 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 (that provides a too optimistic evaluation of the classifier's performance) and the holdout method (non-advisable in unbalanced learning because the scarcity of rare class data prevents their use in both estimating and testing the model).

a. Discussion

As mentioned in Section a, the idea of generating artificial examples similar to the observed sample in order to provide for the class imbalance has been already developed by some authors. 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, most of the proposed 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) and Lee (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, but the link between ROSE and the kernel methods allows us to consider each $\mathbf{H}_j$ as a smoothing matrix and to choose it as the solution of one of the several methods of bandwidth selection proposed in the literature. For a review, see, for example, Wand and Jones (1995).

- 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. As a matter of fact, 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 exactly 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. This entails that the synthetic training set does not even partially overlap the original one, thus reducing the risk that the model overfits the data and giving the opportunity of using the observations for testing the classifier.

- 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 unbalanced data, doubling the size of the rare class may help the learning process only in moderately unbalanced 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 critical point consists of an adequate selection of the smoothing parameters rather than the kernel function, 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 proposed in Lee (1999) is described in Lee (2000), which is 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 the corresponding classifiers are trained. Afterward, the estimated conditional probabilities obtained by each classifier are averaged across the generated samples. It is well-known that, in general, combining several versions of the same classifier aid the improvement of the performance of a single model, although the computational complexity increases. However, in extremely unbalanced 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 according to ROSE, prior to estimating the model, has the beneficial interpretation of building a bagging classifier (Breiman, 1996).

ROSE and the regularization practice proposed by Lee have been further compared through a small simulation study in order to understand if the two methods differ only on paper or if the mentioned differences actually have an impact on the classification. The estimation of a standard classifier without using any remedy for dealing with imbalance has been considered as a benchmark for evaluating the performance of the two methods. Moreover, a bagging version of ROSE has been tested by repeatedly bootstrapping the two classes.

For comparative purposes, 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 considered classifiers (classification trees and logit models) have been estimated on a training

TABLE 2. 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 not result empty); *mult* is the number of noisy replicates for each rare example in Lee (2000); that is, the training set dimension is *mult*·number of minority examples+ number of majority examples. In order to compare the two methods on equal terms the balanced sample generated according to ROSE has the same size. 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 from the two classes according to ROSE.

distribution of data	$(\mathbf{x}, y)$ s. t. $\begin{cases} \mathbf{x} \sim \mathbf{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) & \text{if } y = 0 \\ \mathbf{x} \sim \mathbf{N}_2\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}\right) & \text{if } y = 1 \end{cases}$
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

set while their performance has been evaluated on a test set. Only simulated data have been considered in order to make the sample size and the proportion of minority examples vary.

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 to be chosen in ROSE have been selected as asymptotically optimal for the gaussian distribution.

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 having the same size as Lee have been generated according to ROSE.

The AUC has been chosen as an evaluation metric to measure the performance of the classifiers. Table 2 summarizes the simulation design.

In Tables 3, 4 and 5, the results referring to the use of a 8-nodes tree 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, and when the minority observations amount to only 1% of the training set, the classifier does not even perform better than a random guess.

On the other hand, ROSE allows for a remarkable improvement of the classifier accuracy. The empirical analysis shows that the larger the original training set, the higher the AUC results, besides a tendency towards a depletion in accuracy, when the imbalance increases, is still evident.

As expected, the regularization method proposed by Lee (2000) also aids the improvement of the performance of the classifier. 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. Interesting considerations arise from the comparison between the regularization method described by Lee and ROSE: when both the number of noisy samples and the number of noisy replicates for each rare example are large, ROSE perform (almost) uniformly 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 5) 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.

The use of logit models (instead of a classification tree) generally leads to higher levels of accuracy, but analogous considerations about the comparative behaviour of ROSE and Lee's method may be drawn.

4. Empirical analysis

The current section aims to understand if the good properties of ROSE correspond to good performance of classification when ROSE is used in unbalanced frameworks. In particular, we will first analyze the opportunity to exploit ROSE to evaluate the performance of classification. Then, we will implement some applications of ROSE to real data in order to show its effectiveness in improving the performance of classification in unbalanced learning.

TABLE 3. AUC obtained when training an 8-node classification tree without using any remedy for unbalanced data. Several sample sizes ($n = 250, 1000, 5000$) and proportions of positive examples ($\pi = 0.10, 0.05, 0.025, 0.01$) are considered.

	unbalanced data		
	n		
π	250	1000	5000
0.10	0.67	0.73	0.62
0.05	0.52	0.61	0.51
0.025	0.50	0.55	0.51
0.01	0.50	0.50	0.50

a. Model evaluation by using ROSE

In Section b, it was outlined that creating new artificial examples from an estimate of the conditional densities of the two classes gives the opportunity to exploit the original observed data to test the accuracy of the estimated classification rule. Now, we give an illustration

TABLE 4. AUC obtained when training an 8-node classification tree after balancing the sample by ROSE. n is the original sample size and π is the proportion of rare events, while the balanced training set generated by ROSE has size $mult$·number of positive examples+ number of negative examples.

	ROSE					
	$mult = 2$			$mult = 10$		
	n			n		
π	250	1000	5000	250	1000	5000
0.10	0.79	0.78	0.80	0.78	0.79	0.80
0.05	0.81	0.80	0.82	0.80	0.83	0.81
0.025	0.75	0.77	0.80	0.77	0.78	0.77
0.05	0.72	0.76	0.78	0.72	0.78	0.76

TABLE 5. Each of the 12 subtables reports the AUC obtained when the classification task is performed on a n-sized sample with proportion of positive examples set to π. $mult$ is used to define the actual size of the training set on which the estimation of the classifier is based (see Table 2 for details). On the left, results have been obtained by adopting the regularization method proposed in Lee (2000), with K iterations; on the right the corresponding results obtained by running a bagged version of ROSE are displayed.

		Lee regularization						Bagged ROSE					
		$mult = 2$			$mult = 10$			$mult = 2$			$mult = 10$		
		n			n			n			n		
	π	250	1000	5000	250	1000	5000	250	1000	5000	250	1000	5000
$K = 5$	0.10	0.79	0.79	0.78	0.83	0.84	0.81	0.84	0.86	0.83	0.83	0.85	0.84
	0.05	0.72	0.73	0.78	0.85	0.81	0.82	0.86	0.85	0.85	0.86	0.86	0.85
	0.025	0.65	0.79	0.70	0.79	0.85	0.77	0.80	0.83	0.85	0.82	0.82	0.84
	0.01	0.50	0.52	0.52	0.68	0.85	0.74	0.75	0.82	0.82	0.72	0.82	0.81
$K = 10$	0.10	0.82	0.83	0.80	0.84	0.85	0.82	0.85	0.87	0.85	0.85	0.86	0.85
	0.05	0.81	0.79	0.78	0.86	0.83	0.83	0.87	0.88	0.86	0.86	0.88	0.86
	0.025	0.64	0.83	0.74	0.83	0.86	0.78	0.83	0.85	0.85	0.83	0.85	0.85
	0.01	0.50	0.66	0.64	0.78	0.79	0.75	0.78	0.82	0.82	0.78	0.82	0.82
$K = 20$	0.10	0.83	0.84	0.80	0.84	0.86	0.83	0.85	0.87	0.85	0.85	0.87	0.85
	0.05	0.84	0.81	0.79	0.88	0.84	0.84	0.87	0.87	0.85	0.87	0.87	0.86
	0.025	0.71	0.84	0.77	0.84	0.87	0.79	0.83	0.86	0.85	0.82	0.86	0.85
	0.01	0.50	0.64	0.69	0.83	0.81	0.77	0.78	0.82	0.83	0.78	0.82	0.83

about the soundness of this practice.

Simulations have been conducted by generating data from the two bivariate gaussian densities mentioned in Table 2, each of them corresponding to one class. Compared to the simulation performed in the previous section, the cardinalities of the two classes are not fixed, but the training set is randomly drawn from a mixture of the two distributions, with mixing proportion governing the class imbalance and varying in the set $\{0.5, 0.1, 0.025, 0.01\}$. 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 classifiers' accuracy.

Again, nonparametric classification trees and logit models have been used as 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 1000 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 1000 samples of size 10000 drawn from the same population as the training samples and averaging the resulting AUCs. These 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 Table 6. 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 unbalanced 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 behaviour occurs regardless of the considered classifier, thus making this estimator totally inadequate for use in a context of unbalanced learning.

The practice of testing the accuracy of the classifier on the originally observed data, after training the classifier on synthetic examples generated according to ROSE, appears to be 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 *ceteris paribus*: when the resubstitution method and ROSE are considered, the observed 1000−sized sample is used as a test set and 1000 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.

b. ROSE in practice

Once that we are confident that creation of artificial training examples by ROSE allows us to successfully exploit the original observations to test the classifier, this technique may be adopted in order to finally analyze the ability of ROSE to improve the ability of the classifier in learning from unbalanced data.

To this end, three real data sets have been considered. The first two applications (also used in Lee, 2000) concern medical diagnosis problems and are available from the UCI machine learning repository (Asuncion and Newman, 2007).

The *hypothyroid* data set includes 25 attributes measured on 3163 individuals. Only the five quantitative variables denoted by TSH, T3, TT4, T4U and FTI have been considered in the analysis, and the observations reporting missing values have been discarded. The preprocessed data set includes 2261 negative examples (healthy individuals) and 137 positive cases (patients affected by hypothyroidism), but the distribution of the classes has been further unbalanced by considering only a proportion of 2.5% rare cases, randomly selected from the class of patients affected by hypothyroidism.

The *pima indians* data set gathers 8 characteristics (physical and clinical measurements) of 768 females of Pima Indians, a population in which a high incidence of diabetes has been historically reported. The response variable is the positive or negative result from a diabetes test. Again, the skewness of the class distribution has been made more extreme by including

TABLE 6. 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. The first table refers to the estimation of a classification tree while the second one reports results from the use of a logit model.

classification tree

		50%	10%	1%
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

in the training set all 500 negative instances and a few randomly selected positive instances (the selected rare examples amount to 1% of the whole data set).

The third considered data set has been built by merging data from the Italian *Infocamere* archive and the Business Register, with the aim of discriminating the defaulter and non-defaulter firms. It consists of some vital statistics (e.g. changes of legal status, occurrence of a corporate merger or breakup, number of employees), balance sheet items and financial ratios of all the commercial companies located in a northeastern province of Italy. A data-cleaning stage and a preliminary selection of the most informative variables has been performed on the available data, thus resulting in 11199 cases and 27 attributes. The occurrence of a bankruptcy condition is considered as the default event. This data set is a notable example of classification in the presence of rare classes, with the proportion of defaulter firms amounting to less than 7‰.

A nonparametric decision tree and a logit model have been chosen as classification models.

The classifiers have been trained on 50 balanced ROSE samples generated from each

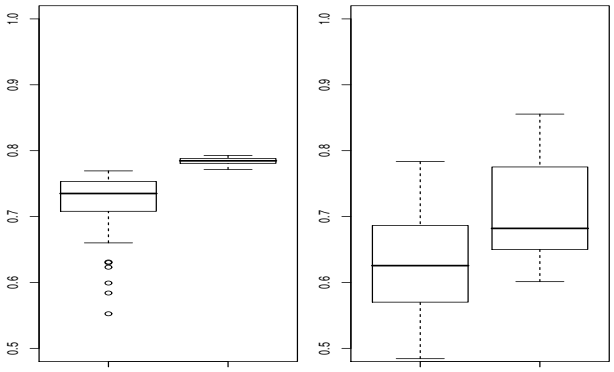

FIG. 1. Distribution of the AUC for the *hypothyroid* data. The left panel refers to the use of a logit model, while the right panel refers to the use of a classification tree. In both panels, the first boxplot displays the AUCs obtained by training the classifier on the smoothed bootstrap unbalanced samples, and the the second boxplot results from training the classifier on ROSE samples.

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 unbalanced 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 Figures 1, 2, 3.

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 lies in the three examples between 0.6 and 0.7, but the variability of the AUC's distributions is high and the inferior whiskers of the plots brush against the value of 0.5. When ROSE is run prior to the tree building, the dispersion of the AUC is not always lower than the corresponding dispersion if the class imbalance is ignored, but the median AUC always exceeds the median AUC resulting from training the tree on unbalanced data. Excellent results are obtained when the *pima* data set is used, since ROSE manages to get an almost perfect prediction.

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 uniformly outperforms predictions based on unbalanced data.

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

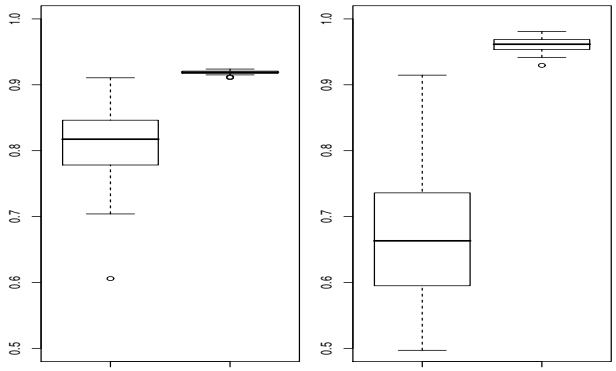

FIG. 2. Distribution of the AUC for the *pima* data. Cf. Figure 1 for further details.

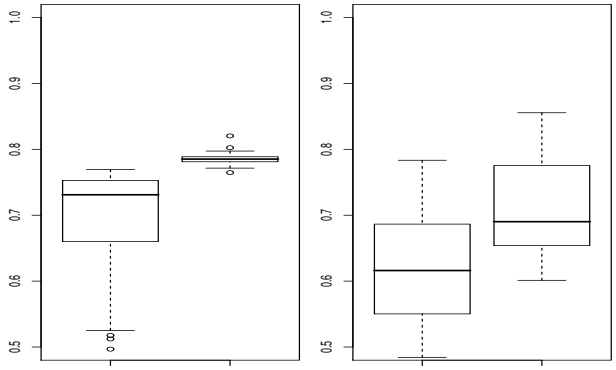

FIG. 3. Distribution of the AUC for the *Infocamere* data. Cf. Figure 1 for further details.

5. Final Remarks

In this work, a comprehensive discussion about the use of unbalanced data in performing binary classification has been provided. In particular, a review of the main causes of the failure of both parametric and nonparametric standard classifiers has been reported, and some new perspectives have been presented about the effects of class imbalance. Indeed, literature dealing with skewed binary classification 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's estimator has been totally ignored. In fact, when the distribution of the classes is skewed, the estimated models perform very poorly but, bad estimates of the classifier's performance may lead to misleading conclusions about the quality of the prediction.

The need to simultaneously deal with both the problems of model estimation and model evaluation has arisen, and a unified and systematic framework has been proposed, based 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. The technique may also be successfully used for an improved estimation of the learner's accuracy and, if one is willing to bear a increased computational complexity, it may be combined with bagging ideas, thus improving the performance of classification even more.

REFERENCES

Akbani, R., Kwek S. and Japkowicz, N. (2004). Applying support vector machines to unbalanced datasets. In: J. F. Boulicaut, F. Esposito, F. Giannotti and D. Pedreschi, Editors, Lecture Notes in Computer Science, ECML: Proceedings of 15th European Conference on Machine Learning, Springer, Pisa, 3201:39-50.

Asuncion, A., Newman, D.J. (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

Barandela, R., Sánchez, J. S., García, V., Rangel, E. (2003). Strategies for learning in class imbalance problems. Pattern Recognition, 36: 849-851.

Bradford, J. P., Kunz, C., Kohavi, R., Brunk C., Brodley, C. E. (1998). Pruning decision trees with misclassification costs. Lecture Notes in Computer Sciences, Springer, Berlin, pp. 131-136.

Breiman, L. (1996). Bagging predictors. Machine Learning, 24:123-140.

Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984). Classification and Regression Trees, Wadsworth International Group, Belmont, California.

Burez, J. and Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36:4626-4636.

Chawla, N. V. (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, N. V., Bowyer, K. W., Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16:321-357.

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

Cieslak, D. and Chawla, N. (2008). Learning decision trees for unbalanced data. Lecture Notes in Computer Science, 5211:241-256.

Cramer, J.S. (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. Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, Pennsylvania. Edited by: Cohen WW, Moore A. New York: ACM Press, pp. 233-240.

Draper B, Carla E. Brodley, C. E., Utgoff, P. E. (1994). Goal-Directed Classification using Linear Machine Decision Trees. IEEE Transactions on Pattern Analysis and Machine Intelligence, 16:888-893.

Drummond, C. and Holte, R. C. (2006). Cost curves: An improved method for visualizing classifier performance. Machine Learning, 65(1):95-130.

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

Eitrich, T., Kless, A., Druska, C., Meyer, W., Grotendorst, J. (2007). Classification of highly unbalanced CYP450 data of drugs using cost sensitive machine learning techniques. Journal of Chemical Information and Modeling, 47(1):92-103.

Estabrooks, A., Taeho, J. and Japkovicz, N. (2004). A multiple resampling Method for Learning form Imbalanced Data Sets. Computational Intelligence, 20(1):18-36.

Fan, W., Stolfo, S., Zhang, J. and Chan, P. (1999). Adacost: Misclassification cost-sensitive boosting. In Proceedings of the 16th International Conference on Machine Learning, pp. 97-105.

Hand, D. (2006). Classifier Technology and the Illusion of Progress. Statistical Science, 21(1):1-14.

Hand, D.J. and Vinciotti, V. (2003). Choosing K for Two-Class Nearest Neighbour Classifiers with Unbalanced Classes, Pattern Recognition Letters, 24: 1555-1562.

He, H. and Garcia, E. A. (2009). Learning from Imbalanced Data, IEEE Transactions on Knowledge and Data Engineering, 21(9).

Guo, H. Viktor, H. L. (2004). Boosting with Data Generation: Improving the Classification of Hard to Learn Examples SIGKDD Explorations, 6(1):30-39.

Japkowicz, N., Stephen, S. (2002). The Class Imbalance Problem: A Systematic Study. Intelligent Data Analysis Journal, vol. 6.

Jo, T. and Japkowicz, N. (2004). Class Imbalances versus Small Disjuncts. SIGKDD Explorations, 6(1):40-49.

Khoshgoftaar, T. M., Golawala, M., and Hulse, J. V. (2007). An Empirical Study of Learning from Imbalanced Data Using Random Forest. Proceedings of the 19th IEEE international Conference on Tools with Artificial intelligence, vol. 2.

King, E. N., and Ryan, T. P. (2002). A Preliminary Investigation of Maximum Likelihood Logistic Regression versus Exact Logistic Regression. The American Statistician, 56:163-170.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. Political Analysis, 9:137-163.

Kotsiantis, S., Kanellopoulos, D., Pintelas, P. (2006). Handling imbalanced datasets:a review. GETS International Transactions on Computer Science and Engineering, 30.

Kukar M. and Kononenko, I. (1998). Cost-Sensitive Learning with Neural Networks, Proc. 13th European Conf. Artificial Intelligence, pp. 445-449.

Kubat, M. and Matwin, S. (1997). Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. Proceedings of the 14th International Conference on Machine Learning, ICML, Nashville, TN, U.S.A, pp. 179-186.

Lawrence, S., Burns, I., Back, A. D., Tsoi A. C. and Lee G. C. (1998). Neural Network Classification and Prior Class Probabilities, State-of-the-Art Surveys, Series Tricks of the Trade, Lecture Notes in Computer Science. G. Orr, K-R. Mueller, R. Caruana (Eds), Springer-Verlag. pp. 299-314.

Lee, S. (2000). Noisy replication in skewed binary classification. Computational Statistics and Data Analysis, 34:165-191.

Lee, S. (1999). Regularization in skewed binary classification. Computational Statistics, 14:277-292.

Lin, Y., Lee, Y., Wahba, G. (2002). Support Vector Machines for Classification in Nonstandard Situations, Machine Learning, 46:191-202

Liu, Y., Chawla, N. V., Harper, M. P., Shriberg, E. and Stolcke, A. (2006). A study in machine learning from imbalanced data for sentence boundary detection in speech. Computer Speech & Language, 20:468-494.

Mazurowski M. A., (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance, Neural Networks, 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, NY, USA, pp. 69-77.

Mease, D., Wyner, A. and Buja, A. (2007) Boosted classification trees and class probability-quantile estimation. Journal of Machine Learning Research, 8:409-439.

Mehta, C. R., and Patel, N. R. (1995) Exact Logistic Regression: Theory and Examples, Statistics in Medicine, 14:2143-2160.

Owen, A. B. (2007). Infinitely imbalanced logistic regression. Journal of Machine Learning Research, 8:761-773.

Riddle, P., Segal, R., Etzioni, O. (1994). Representation Design and Brute-Force Induction in a Boeing Manufacturing Domain. Applied Artificial Intelligence, 8:125-147.

Schiavo, R. A., Hand, D. J. (2000) Ten more years of error rate research. International Statistical Review, 68(3):295-310.

Silverman, B. W. (1986). Density estimation for statistics and data analysis. New York: Chapman and Hall.

Sun, Y., Kamel, M.S., Wong, A.K.C., Wang, Y. (2007) Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition, 40(12):3358-3378.

Veropoulos, K., Campbell, C., Cristianini, N. (1999) Controlling the sensitivity of support vector machines. Proceedings of International Joint Conference on Artificial Intelligence, pp. 55-60.

Wu, G., Chang, E. Y. (2005) KBA: Kernel Boundary Alignment Considering Imbalanced Data Distribution, IEEE Transactions on Knowledge and Data Engineering, 17(6):786-795.

Xie, J. and Qiu, Z. (2007). The effect of imbalanced data sets on LDA: A theoretical and empirical analysis, Pattern Recognition, 40(2):557-562.

Ting, K. M. (2002) An Instance-Weighting Method to Induce Cost-Sensitive Trees. IEEE Trans. Knowl. Data Eng, 14(3):659-665.

Thomas, J. Jouve, P. ,and Nicoloyannis, N. (2006) Optimisation and evaluation of random forests for imbalanced datasets. Lecture Notes in Computer Science, Springer, 4203:622-631.

Wand, M. P. and Jones, M. C. (1995). Kernel Smoothing. London: Chapman and Hall.

Wehberg, S. and Schumacher, M. (2004) A Comparison of Nonparametric Error Rate Estimation Methods in Classification Problems. Biometrical Journal, 46(1):35-47.

Weiss, G. M. (2004). Mining with rarity: a unifying framework, ACM SIGKDD Explorations Newsletter, 6(1).

Weiss, G. M., 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.

Xue, J.H. and Titterington, D.M. (2008). Do unbalanced data have a negative effect on LDA? Pattern Recognition, 41(5):1575-1588.

Yen, S. and 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 Transactions on Knowledge and Data Engineering, 18(1):63-77.