On the Statistical Consistency of Algorithms for Binary Classification under Class Imbalance

Aditya Krishna Menon AKMENON@UCSD.EDU
University of California, San Diego, La Jolla CA 92093, USA

Harikrishna Narasimhan HARIKRISHNA@CSA.IISC.ERNET.IN
Shivani Agarwal SHIVANI@CSA.IISC.ERNET.IN
Indian Institute of Science, Bangalore 560012, India

Sanjay Chawla SANJAY.CHAWLA@SYDNEY.EDU.AU
University of Sydney and NICTA, Sydney, Australia

Abstract

Class imbalance situations, where one class is rare compared to the other, arise frequently in machine learning applications. It is well known that the usual misclassification error is ill-suited for measuring performance in such settings. A wide range of performance measures have been proposed for this problem. However, despite the large number of studies on this problem, little is understood about the statistical consistency of the algorithms proposed with respect to the performance measures of interest. In this paper, we study consistency with respect to one such performance measure, namely the arithmetic mean of the true positive and true negative rates (AM), and establish that some practically popular approaches, such as applying an empirically determined threshold to a suitable class probability estimate or performing an empirically balanced form of risk minimization, are in fact consistent with respect to the AM (under mild conditions on the underlying distribution). Experimental results confirm our consistency theorems.

1. Introduction

Classification problems with class imbalance – where one class is rare compared to another – arise in several machine learning applications, ranging from medical diagnosis and text retrieval to credit risk prediction and fraud detection. Due to their practical importance, such class imbalance settings have been widely studied in several fields, including machine learning, data mining, artificial intelligence, and several others (Cardie & Howe, 1997; Kubat & Matwin, 1997; Japkowicz, 2000; Elkan, 2001; Japkowicz & Stephen, 2002; Chawla et al., 2002; 2003; Zadrozny et al., 2003; Chawla et al., 2004; Drummond & Holte, 2005; 2006; Van Hulse et al., 2007; Chawla et al., 2008; Qiao & Liu, 2009; Gu et al., 2009; He & Garcia, 2009; Liu & Chawla, 2011; Wallace et al., 2011).

The usual misclassification error is ill-suited as a performance measure in class imbalance settings, since a default classifier predicting the majority class does well under this measure. A variety of performance measures have been proposed for evaluating binary classifiers in such settings; these include for example the arithmetic, geometric, and harmonic means of the true positive and true negative rates, which attempt to balance the errors on the two classes, and related measures based on the recall (true positive rate) and precision (see Table 1). Several algorithmic approaches have also been proposed, for example under-sampling the majority class, over-sampling the minority class, changing the decision threshold of a score-based classifier, and modifying algorithms to incorporate different weights for errors on positive and negative examples.

Despite the large number of studies on the class imbalance problem, surprisingly little is understood about the statistical consistency of the algorithms proposed with respect to the performance measures of interest, i.e. about whether the algorithms converge to the optimal value of the performance measure in the large

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

Table 1. Performance measures that have been used to evaluate binary classifiers in class imbalance settings. Here TPR, TNR, FPR, and FNR denote the true positive, true negative, false positive, and false negative rates, respectively; and Prec denotes precision. Note that AUC-ROC/AUC-PR apply not to a classifier but to a *scoring function* (or more generally, to a family of classifiers; a scoring function yields a family of classifiers via different thresholds); we include these measures here since some studies have used these with the aim of evaluating binary classification methods under class imbalance.

Measure	Definition	References
A-Mean (AM)	$(\text{TPR} + \text{TNR})/2$	(Chan & Stolfo, 1998; Powers et al., 2005; Gu et al., 2009) KDD Cup 2001 challenge (Cheng et al., 2002)
G-Mean (GM)	$\sqrt{\text{TPR} \cdot \text{TNR}}$	(Kubat & Matwin, 1997; Daskalaki et al., 2006)
H-Mean (HM)	$2/(\frac{1}{\text{TPR}} + \frac{1}{\text{TNR}})$	(Kennedy et al., 2009)
Q-Mean (QM)	$1 - ((\text{FPR})^2 + (\text{FNR})^2)/2$	(Lawrence et al., 1998)
F_1	$2/(\frac{1}{\text{Prec}} + \frac{1}{\text{TPR}})$	(Lewis & Gale, 1994; Gu et al., 2009)
G-TP/PR	$\sqrt{\text{TPR} \cdot \text{Prec}}$	(Daskalaki et al., 2006)
AUC-ROC	Area under ROC curve	(Ling et al., 1998)
AUC-PR	Area under precision-recall curve	(Davis & Goadrich, 2006; Liu & Chawla, 2011)

sample limit. In this paper, we study this question for one such performance measure that is widely used in class imbalance settings, namely the arithmetic mean of the true positive and true negative rates (AM). The usual Bayes optimal classifier that minimizes the classification error rate is not optimal for this measure, and therefore standard binary classification algorithms that are designed to converge to the Bayes error are not consistent with respect to this measure. We show consistency with respect to the AM measure (under mild conditions on the underlying distribution) of two simple families of algorithms that have been used in class imbalance settings in practice: (1) algorithms that apply a suitable threshold (determined from the empirical class ratio) to a class probability estimate obtained by minimizing an appropriate strongly proper loss, and (2) algorithms that minimize a suitably weighted form of an appropriate classification-calibrated loss (with the weights determined from the empirical class ratio).

Our results build on several recent tools that have been developed for studying consistency of learning algorithms: regret bounds for standard binary classification using classification-calibrated losses (Zhang, 2004; Bartlett et al., 2006), proper and strongly proper losses (Reid & Williamson, 2009; 2010; Agarwal, 2013), balanced losses that have been used recently to understand consistency in ranking problems (Kotlowski et al., 2011), and regret bounds for cost-sensitive classification (Scott, 2012). In addition, a key tool we introduce is a decomposition lemma that allows us to reduce the problem of analyzing the AM regret for class imbalance settings to analyzing an *empirical* cost-sensitive regret, in which the cost parameter is determined from the empirical class ratio. For each of the above two families of algorithms, we then show that under suitable conditions, this empirical cost-sensitive regret converges in probability to zero, thus establishing consistency with respect to the AM measure.

The paper is organized as follows. Section 2 contains preliminaries and background; Section 3 contains our decomposition lemma. Sections 4–5 give consistency results for the above two families of algorithms, respectively. Section 6 contains our experimental results.

2. Preliminaries and Background

2.1. Problem Setup and Notation

Let $\mathcal{X}$ be any instance space. Given a training sample $S = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \{\pm 1\})^n$, the goal is to learn a binary classifier $h_S : \mathcal{X} \to \{\pm 1\}$ to classify new instances in $\mathcal{X}$. Assume all examples (both training examples and future test examples) are drawn iid according to some unknown probability distribution D on $\mathcal{X} \times \{\pm 1\}$. Let $\eta(x) = \mathbf{P}(y = 1|x)$, and let $p = \mathbf{P}(y = 1)$ (both under D).We shall assume $p \in (0, 1)$. In the class imbalance setting, p departs significantly from $\frac{1}{2}$; by convention, we assume the positive class is rare, so that p is small.

For any candidate classifier $h : \mathcal{X} \to \{\pm 1\}$, we can define the true positive rate (TPR), true negative rate (TNR), false positive rate (FPR), false negative rate (FNR), and precision (Prec) of h w.r.t. D as follows:

$$\begin{aligned}
\text{TPR}_D[h] &= \mathbf{P}(h(x) = 1 \,|\, y = 1) \\
\text{TNR}_D[h] &= \mathbf{P}(h(x) = -1 \,|\, y = -1) \\
\text{FPR}_D[h] &= \mathbf{P}(h(x) = 1 \,|\, y = -1) \;=\; 1 - \text{TNR}_D[h] \\
\text{FNR}_D[h] &= \mathbf{P}(h(x) = -1 \,|\, y = 1) \;=\; 1 - \text{TPR}_D[h] \\
\text{Prec}_D[h] &= \mathbf{P}(y = 1 \,|\, h(x) = 1)
\end{aligned}$$

As noted in Section 1, a variety of performance measures that combine the above quantities and seek to balance errors on the two classes have been proposed for classification with class imbalance (see Table 1). In this work, we focus on the arithmetic mean of the TPR and TNR (AM):

$$\mathrm{AM}_D[h] \;=\; \frac{\mathrm{TPR}_D[h] + \mathrm{TNR}_D[h]}{2}\,.$$

In particular, we would like to find a classifier that has AM performance close to the optimal:

$$\mathrm{AM}_D^* = \sup_{h:\mathcal{X}\to\{\pm 1\}} \mathrm{AM}_D[h]\,.$$

More precisely, define the AM-*regret* of h as

$$\mathrm{regret}_D^{\mathrm{AM}}[h] = \mathrm{AM}_D^* - \mathrm{AM}_D[h]\,.$$

Then we would like an algorithm to be AM-*consistent*, i.e. we would like the AM-regret of the learned classifier h_S to converge in probability to zero:[1]

$$\mathrm{regret}_D^{\mathrm{AM}}[h_S] \xrightarrow{\mathrm{P}} 0\,.$$

Next we recall various notions related to loss functions that will be used in our study.

2.2. Loss Functions

A binary loss function on a prediction space $\widehat{\mathcal{Y}} \subseteq \bar{\mathbb{R}}$ is a function $\ell : \{\pm 1\} \times \widehat{\mathcal{Y}} \to \bar{\mathbb{R}}_+$ that defines a penalty $\ell(y, \widehat{y})$ incurred on predicting $\widehat{y} \in \widehat{\mathcal{Y}}$ when the true label is $y \in \{\pm 1\}$ (here $\bar{\mathbb{R}} = [-\infty, \infty]$, $\bar{\mathbb{R}}_+ = [0, \infty]$). The ℓ-error of a function $f : \mathcal{X} \to \widehat{\mathcal{Y}}$ w.r.t. D is then

$$\mathrm{er}_D^{\ell}[f] = \mathbf{E}_{(x,y)\sim D}\big[\ell(y, f(x))\big]\,.$$

The optimal ℓ-error w.r.t. D is

$$\mathrm{er}_D^{\ell,*} = \inf_{f:\mathcal{X}\to\widehat{\mathcal{Y}}} \mathrm{er}_D^{\ell}[f]\,,$$

and the ℓ-regret of f w.r.t. D is

$$\mathrm{regret}_D^{\ell}[f] = \mathrm{er}_D^{\ell}[f] - \mathrm{er}_D^{\ell,*}\,.$$

As an example, for $\widehat{\mathcal{Y}} = \{\pm 1\}$, the familiar 0-1 loss $\ell_{0\text{-}1} : \{\pm 1\} \times \{\pm 1\} \to \bar{\mathbb{R}}_+$ takes the form

$$\ell_{0\text{-}1}(y, \widehat{y}) = \mathbf{1}(\widehat{y} \neq y)\,,$$

where $\mathbf{1}(\cdot)$ denotes the indicator function with value 1 if its argument is true and 0 otherwise, and the $\ell_{0\text{-}1}$-error of $h : \mathcal{X} \to \{\pm 1\}$ w.r.t. D takes the form

$$\mathrm{er}_D^{0\text{-}1}[h] = \mathbf{E}_{(x,y)\sim D}\big[\mathbf{1}(h(x) \neq y)\big]\,.$$

In this case the optimal $\ell_{0\text{-}1}$-error $\mathrm{er}_D^{0\text{-}1,*}$ is the usual Bayes classification error.

For any loss $\ell : \{\pm 1\} \times \widehat{\mathcal{Y}} \to \bar{\mathbb{R}}_+$, the *conditional ℓ-risk* $L_\ell : [0,1] \times \widehat{\mathcal{Y}} \to \bar{\mathbb{R}}_+$ is defined as[2]

$$L_\ell(\eta, \widehat{y}) = \eta\, \ell(1, \widehat{y}) + (1 - \eta)\, \ell(-1, \widehat{y})\,,$$

and the *conditional Bayes ℓ-risk* $H_\ell : [0,1] \to \bar{\mathbb{R}}_+$ is defined as

$$H_\ell(\eta) = \inf_{\widehat{y}\in\widehat{\mathcal{Y}}} L_\ell(\eta, \widehat{y})\,.$$

Cost-Sensitive Losses. For any base loss $\ell : \{\pm 1\} \times \widehat{\mathcal{Y}} \to \bar{\mathbb{R}}_+$ and $c \in (0,1)$, the *cost-sensitive* ℓ-loss with cost parameter c, $\ell^{(c)} : \{\pm 1\} \times \widehat{\mathcal{Y}} \to \bar{\mathbb{R}}_+$, is defined as

$$\ell^{(c)}(y, \widehat{y}) = \Big((1 - c)\,\mathbf{1}(y = 1) + c\,\mathbf{1}(y = -1)\Big) \cdot \ell(y, \widehat{y})\,.$$

Note that for the 0-1 loss $\ell_{0\text{-}1}$, the cost-sensitive loss $\ell_{0\text{-}1}^{(c)}$ with cost parameter c simply assigns a cost of c to false positives and $1 - c$ to false negatives:

$$\ell_{0\text{-}1}^{(c)}(y, \widehat{y}) = (1-c)\,\mathbf{1}(y = 1, \widehat{y} = -1) + c\,\mathbf{1}(y = -1, \widehat{y} = 1)\,.$$

It is well known that for any $\eta \in [0,1]$, $L_{0\text{-}1}^{(c)}(\eta, \widehat{y})$ is minimized by $\widehat{y}^* = \mathrm{sign}(\eta - c)$, and therefore an optimal classifier w.r.t. the $\ell_{0\text{-}1}^{(c)}$-error is given by $h_c^*(x) = \mathrm{sign}(\eta(x) - c)$ (Elkan, 2001).[3]

Balanced Losses. For any base loss $\ell : \{\pm 1\} \times \widehat{\mathcal{Y}} \to \bar{\mathbb{R}}_+$ and distribution D with $p = \mathbf{P}(y = 1) \in (0,1)$, the *balanced* ℓ-loss $\ell^{\mathrm{bal}} : \{\pm 1\} \times \widehat{\mathcal{Y}} \to \bar{\mathbb{R}}_+$ is defined as

$$\ell^{\mathrm{bal}}(y, \widehat{y}) = \Big(\frac{\mathbf{1}(y = 1)}{2p} + \frac{\mathbf{1}(y = -1)}{2(1 - p)}\Big) \cdot \ell(y, \widehat{y})\,.$$

Note that a balanced loss depends on the underlying distribution D via p and therefore typically cannot be evaluated directly; however it is a useful analytical tool that has been used recently to analyze consistency of ranking algorithms in (Kotlowski et al., 2011), and will be useful for our purposes as well.

Classification-Calibrated Losses. Let $c \in (0,1)$. A loss $\ell : \{\pm 1\} \times \bar{\mathbb{R}} \to \bar{\mathbb{R}}_+$ is said to be *classification-calibrated at c* (Bartlett et al., 2006; Reid & Williamson, 2010; Scott, 2012) if $\forall \eta \in [0,1], \eta \neq c$,

$$\inf_{f\in\bar{\mathbb{R}}: f(\eta - c) \leq 0} L_\ell(\eta, f) \;>\; H_\ell(\eta)\,.$$

This condition ensures that $\forall \eta \in [0,1]$, if $f^* \in \bar{\mathbb{R}}$ is a minimizer of $L_\ell(\eta, f)$, then $\widehat{y}^* = \mathrm{sign}(f^*)$ is a minimizer of $L_{0\text{-}1}^{(c)}(\eta, \widehat{y})$ (see cost-sensitive losses above). For $c = \frac{1}{2}$, this ensures if f^* is a minimizer of $L_\ell(\eta, f)$, then $\widehat{y}^* = \mathrm{sign}(f^*)$ is a minimizer of $L_{0\text{-}1}(\eta, \widehat{y})$.

[1] Recall $\phi(S)$ converges in probability to $a \in \mathbb{R}$, written $\phi(S) \xrightarrow{\mathrm{P}} a$, if $\forall \epsilon > 0$, $\mathbf{P}_{S\sim D^n}(|\phi(S) - a| > \epsilon) \to 0$ as $n \to \infty$.

[2] We overload η here to denote a number in $[0,1]$ rather than a function; the usage should be clear from context.

[3] Here $\mathrm{sign}(z) = 1$ if $z > 0$ and -1 otherwise.

Proper and Strongly Proper Losses. Proper losses in their basic form are defined on $\widehat{\mathcal{Y}} = [0,1]$ and facilitate class probability estimation. A loss $\ell : \{\pm 1\} \times [0,1] \to \bar{\mathbb{R}}_+$ is said to be *proper* (Reid & Williamson, 2009; 2010) if $\forall \eta \in [0,1]$,

$$L_\ell(\eta, \eta) = H_\ell(\eta)\,.$$

This condition ensures that $\forall \eta \in [0,1]$, the set of minimizers of $L_\ell(\eta, \widehat{\eta})$ (over $\widehat{\eta}$) includes the right value η. A proper loss is said to be *strongly proper* (Agarwal, 2013) if $\exists \kappa > 0$ such that $\forall \eta, \widehat{\eta} \in [0,1]$,

$$L_\ell(\eta, \widehat{\eta}) - H_\ell(\eta) \geq \kappa(\widehat{\eta} - \eta)^2\,.$$

This condition ensures that $\forall \eta \in [0,1]$, η is the unique minimizer of $L_\ell(\eta, \widehat{\eta})$ (over $\widehat{\eta}$), and moreover $H_\ell(\eta) = L_\ell(\eta, \eta)$ is well separated from $L_\ell(\eta, \widehat{\eta})$ for $\widehat{\eta} \neq \eta$.

A loss $\ell : \{\pm 1\} \times \widehat{\mathcal{Y}} \to \bar{\mathbb{R}}_+$ on a general prediction space $\widehat{\mathcal{Y}} \subseteq \mathbb{R}$ is said to be *(strongly) proper composite* (Reid & Williamson, 2010; Agarwal, 2013) if $\exists$ a (strongly) proper loss $\gamma : \{\pm 1\} \times [0,1] \to \bar{\mathbb{R}}_+$ and invertible 'link' function $\psi : [0,1] \to \widehat{\mathcal{Y}}$ such that $\forall y \in \{\pm 1\}, \widehat{y} \in \widehat{\mathcal{Y}}$,

$$\ell(y, \widehat{y}) = \gamma(y, \psi^{-1}(\widehat{y}))\,.$$

3. Decomposition Lemma

We now prove a key decomposition lemma that will allow us to reduce the problem of analyzing the AM-regret of a classifier h_S learned from a sample S to the problem of analyzing a certain empirical cost-sensitive regret derived from S (for distributions D satisfying a mild assumption, namely Assumption A below). We start with the following simple equivalence between the AM measure and (one minus) the balanced 0-1 error:

Proposition 1. *For any $h : \mathcal{X} \to \{\pm 1\}$,*

$$\mathrm{AM}_D[h] = 1 - \mathrm{er}_D^{\text{0-1,bal}}[h]\,.$$

Proof. We have,

$$\begin{aligned}
\mathrm{er}_D^{\text{0-1,bal}}[h] &= \mathbf{E}_{(x,y)\sim D}\left[\ell_{\text{0-1}}^{\text{bal}}(y, h(x))\right] \\
&= \mathbf{E}_{(x,y)\sim D}\left[\left(\frac{\mathbf{1}(y=1)}{2p} + \frac{\mathbf{1}(y=-1)}{2(1-p)}\right) \cdot \mathbf{1}(h(x) \neq y)\right] \\
&= \frac{\mathbf{P}(y=1, h(x)=-1)}{2p} + \frac{\mathbf{P}(y=-1, h(x)=1)}{2(1-p)} \\
&= \frac{\mathbf{P}(h(x)=-1 \,|\, y=1)}{2} + \frac{\mathbf{P}(h(x)=1 \,|\, y=-1)}{2} \\
&= \frac{\mathrm{FNR}_D[h] + \mathrm{FPR}_D[h]}{2} \\
&= 1 - \mathrm{AM}_D[h]\,.
\end{aligned}$$

□

In particular, the above result implies that the AM-regret is equal to the balanced 0-1 regret:

$$\mathrm{regret}_D^{\mathrm{AM}}[h] = \mathrm{er}_D^{\text{0-1,bal}}[h] - \mathrm{er}_D^{\text{0-1,bal},*} \tag{1}$$

The lemma below will need the following assumption:

Assumption A. *We say a probability distribution D on $\mathcal{X} \times \{\pm 1\}$ with $\eta(x) = \mathbf{P}(y=1|x)$ and $p = \mathbf{P}(y=1)$ satisfies assumption A if the cumulative distribution functions of the random variable $\eta(x)$ conditioned on $y=1$ and on $y=-1$, $F_{\eta(x)|y=1}(z) = \mathbf{P}(\eta(x) \leq z \mid y=1)$ and $F_{\eta(x)|y=-1}(z) = \mathbf{P}(\eta(x) \leq z \mid y=-1)$, are continuous at $z = p$.*

We note that the above assumption is quite mild, in that it holds for any distribution D for which the random variable $\eta(x)$ conditioned on $y=1$ and on $y=-1$ is continuous, and also for any distribution D for which $\eta(x)$ conditioned on $y=1$ and on $y=-1$ is mixed or discrete as long as p is not a point of discontinuity.

The decomposition lemma below requires an empirical estimator $\widehat{p}_S$ of $p = \mathbf{P}(y=1)$ that lies in $(0,1)$ and converges in probability to p; while we can use any such estimator, for concreteness, we will consider the following simple estimator in our study:

$$\widehat{p}_S = \begin{cases} \frac{n_S^+}{n} & \text{if } 0 < n_S^+ < n \\ \frac{1}{n} & \text{if } n_S^+ = 0 \\ \frac{n-1}{n} & \text{if } n_S^+ = n \end{cases} \quad ; \quad n_S^+ = \sum_{i=1}^n \mathbf{1}(y_i = 1)\,. \tag{2}$$

It is easy to verify that this estimator satisfies $\widehat{p}_S \in (0,1)$ and $\widehat{p}_S \xrightarrow{\mathrm{P}} p$.

Lemma 2 (Decomposition Lemma). *Let D be a probability distribution on $\mathcal{X} \times \{\pm 1\}$ satisfying Assumption A. Let $h_S : \mathcal{X} \to \{\pm 1\}$ denote the classifier learned by an algorithm from training sample S, and let $\widehat{p}_S$ denote any estimator of $p = \mathbf{P}(y=1)$ satisfying $\widehat{p}_S \in (0,1)$ and $\widehat{p}_S \xrightarrow{P} p$. If the empirical cost-sensitive 0-1 regret of h_S with cost parameter $c = \widehat{p}_S$ satisfies*

$$\mathrm{regret}_D^{\text{0-1},(\widehat{p}_S)}[h_S] \xrightarrow{P} 0\,,$$

then

$$\mathrm{regret}_D^{\mathrm{AM}}[h_S] \xrightarrow{P} 0\,.$$

Proof. From Eq. (1), for any training sample S, we may express the AM-regret of h_S as:

$$\begin{aligned}
\mathrm{regret}_D^{\mathrm{AM}}[h_S] &= \mathrm{er}_D^{\text{0-1,bal}}[h_S] - \mathrm{er}_D^{\text{0-1,bal},*} \\
&= \left(\mathrm{er}_D^{\text{0-1,bal}}[h_S] - \frac{1}{2\widehat{p}_S(1-\widehat{p}_S)}\mathrm{er}_D^{\text{0-1},(\widehat{p}_S)}[h_S]\right) \\
&\quad + \frac{1}{2\widehat{p}_S(1-\widehat{p}_S)}\left(\mathrm{er}_D^{\text{0-1},(\widehat{p}_S)}[h_S] - \mathrm{er}_D^{\text{0-1},(\widehat{p}_S),*}\right) \\
&\quad + \left(\frac{1}{2\widehat{p}_S(1-\widehat{p}_S)}\mathrm{er}_D^{\text{0-1},(\widehat{p}_S),*} - \mathrm{er}_D^{\text{0-1,bal},*}\right)\,.
\end{aligned}$$

If $\text{regret}_D^{\text{0-1},(\widehat{p}_S)}[h_S] \xrightarrow{\text{P}} 0$, then the second term in the above decomposition converges in probability to $1/(2p(1-p)) \cdot 0 = 0$. We will show that under the given conditions, the other two terms also converge in probability to zero, thereby establishing the result.

For the first term, we have

$$\begin{aligned}
&\text{er}_D^{\text{0-1,bal}}[h_S] - \frac{1}{2\widehat{p}_S(1-\widehat{p}_S)}\text{er}_D^{\text{0-1},(\widehat{p}_S)}[h_S] \\
&= \mathbf{E}_{(x,y)\sim D}\Big[\Big(\frac{1}{2p} - \frac{1}{2\widehat{p}_S}\Big)\mathbf{1}\big(y=1, h_S(x)=-1\big) \\
&\quad + \Big(\frac{1}{2(1-p)} - \frac{1}{2(1-\widehat{p}_S)}\Big)\mathbf{1}\big(y=-1, h_S(x)=1\big)\Big] \\
&= \frac{1}{2}\Big(1 - \frac{p}{\widehat{p}_S}\Big) \cdot \text{FNR}_D[h_S] \\
&\qquad + \frac{1}{2}\Big(1 - \frac{1-p}{1-\widehat{p}_S}\Big) \cdot \text{FPR}_D[h_S] \\
&\xrightarrow{\text{P}} 0\,,
\end{aligned}$$

since $\widehat{p}_S \xrightarrow{\text{P}} p$ and since $0 \leq \text{FNR}_D[h_S], \text{FPR}_D[h_S] \leq 1$.

Now, let $h^*_{\widehat{p}_S}(x) = \text{sign}(\eta(x) - \widehat{p}_S)$ and let $h^*_p(x) = \text{sign}(\eta(x) - p)$. Then for the third term in the decomposition, it can be seen that

$$\begin{aligned}
&\frac{1}{2\widehat{p}_S(1-\widehat{p}_S)}\text{er}_D^{\text{0-1},(\widehat{p}_S),*} - \text{er}_D^{\text{0-1,bal},*} \\
&= \frac{1}{2\widehat{p}_S(1-\widehat{p}_S)}\text{er}_D^{\text{0-1},(\widehat{p}_S)}[h^*_{\widehat{p}_S}] - \text{er}_D^{\text{0-1,bal}}[h^*_p] \\
&= \frac{1}{2}\Big(\frac{p}{\widehat{p}_S}\text{FNR}_D[h^*_{\widehat{p}_S}] - \text{FNR}_D[h^*_p]\Big) \\
&\quad + \frac{1}{2}\Big(\frac{1-p}{1-\widehat{p}_S}\text{FPR}_D[h^*_{\widehat{p}_S}] - \text{FPR}_D[h^*_p]\Big)\,.
\end{aligned}$$

Now, $\text{FNR}_D[h^*_{\widehat{p}_S}] = \mathbf{P}\big(\eta(x) \leq \widehat{p}_S \,|\, y=1\big) \xrightarrow{\text{P}} \mathbf{P}\big(\eta(x) \leq p \,|\, y=1\big) = \text{FNR}_D[h^*_p]$, since $\widehat{p}_S \xrightarrow{\text{P}} p$ and by continuity of the cumulative distribution function of $\eta(x)$ given $y=1$ at p (Assumption A). Similarly, $\text{FPR}_D[h^*_{\widehat{p}_S}] \xrightarrow{\text{P}} \text{FPR}_D[h^*_p]$. Combining with the above and the fact that $\widehat{p}_S \xrightarrow{\text{P}} p$ then yields that the third term above also converges in probability to zero. □

Thus, to show AM consistency of an algorithm (for distributions satisfying Assumption A), it suffices to show the empirical cost-sensitive regret in the above lemma converges to zero. Note that this cannot be achieved by a direct application of results for cost-sensitive learning, since the costs there have to be fixed in advance. In the next two sections we show AM consistency for two families of algorithms that have often been used in class imbalance settings in practice.

Algorithm 1 Plug-in with Empirical Threshold

1: **Input:** $S = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \{\pm 1\})^n$
2: **Select:** (a) Proper (composite) loss $\ell : \{\pm 1\} \times \bar{\mathbb{R}} \to \bar{\mathbb{R}}_+$, with link function $\psi : [0,1] \to \bar{\mathbb{R}}$; (b) RKHS $\mathcal{F}_K$ with positive definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$; (c) regularization parameter $\lambda_n > 0$
3: $f_S \in \text{argmin}_{f \in \mathcal{F}_K} \big\{\frac{1}{n}\sum_{i=1}^n \ell(y_i, f(x_i)) + \lambda_n \|f\|_K^2\big\}$
4: $\widehat{\eta}_S = \psi^{-1} \circ f_S$
5: $\widehat{p}_S =$ (as in Eq. (2))
6: **Output:** Classifier $h_S(x) = \text{sign}(\widehat{\eta}_S(x) - \widehat{p}_S)$

4. Consistency of Plug-in Rules with an Empirical Threshold

From Proposition 1, it follows that an optimal classifier w.r.t. the AM measure has the form

$$h^*_p(x) = \text{sign}(\eta(x) - p)\,.$$

Therefore if p were known, one could use a class probability estimate $\widehat{\eta}_S(x)$, such as one obtained by suitably regularized empirical risk minimization (ERM) using a proper loss, and construct a plug-in classifier $\text{sign}(\widehat{\eta}_S(x) - p)$. In the absence of knowledge of p, a natural approach is to use an estimate $\widehat{p}_S$, yielding a plug-in classifier with an *empirical* threshold, $\text{sign}(\widehat{\eta}_S(x) - \widehat{p}_S)$; see Algorithm 1 for a prototype using proper losses in a reproducing kernel Hilbert space (RKHS). The following establishes general conditions under which such algorithms are AM-consistent:

Theorem 3. *Let D be a probability distribution on $\mathcal{X} \times \{\pm 1\}$ satisfying Assumption A. Let $\widehat{p}_S$ denote any estimator of $p = \mathbf{P}(y=1)$ satisfying $\widehat{p}_S \in (0,1)$ and $\widehat{p}_S \xrightarrow{P} p$. Let $\widehat{\eta}_S : \mathcal{X} \to [0,1]$ denote any class probability estimator satisfying $\mathbf{E}_x\big[|\widehat{\eta}_S(x) - \eta(x)|^r\big] \xrightarrow{P} 0$ for some $r \geq 1$, and let $h_S(x) = \text{sign}(\widehat{\eta}_S(x) - \widehat{p}_S)$. Then*

$$\text{regret}_D^{\text{AM}}[h_S] \xrightarrow{P} 0\,.$$

The proof makes use of the following lemma:

Lemma 4. *Let $c \in (0,1)$ and $r \geq 1$. For any $\widehat{\eta} : \mathcal{X} \to [0,1]$ and $h(x) = \text{sign}(\widehat{\eta}(x) - c)$,*

$$\text{regret}_D^{\text{0-1},(c)}[h] \;\leq\; \big(\mathbf{E}_x\big[|\widehat{\eta}(x) - \eta(x)|^r\big]\big)^{1/r}\,.$$

Lemma 4 follows directly from a result of (Scott, 2012); see the supplementary material for details.

Proof of Theorem 3. By Lemma 4, we have

$$\begin{aligned}
\text{regret}_D^{\text{0-1},(\widehat{p}_S)}[h_S] &\leq \big(\mathbf{E}_x\big[|\widehat{\eta}_S(x) - \eta(x)|^r\big]\big)^{1/r} \\
&\xrightarrow{\text{P}} 0 \quad \text{(by assumption)}\,.
\end{aligned}$$

The result follows from Lemma 2. □

Table 2. Examples of loss functions satisfying conditions of Theorems 5 and 6. Here $z_+ = \max(0, z)$.

Loss	$\ell(y, f)$	Theorem 5	Theorem 6
Logistic	$\ln(1 + e^{-yf})$	✓	✓
Exponential	e^{-yf}	✓	✓
Square	$(1 - yf)^2$	✓	✓
Sq. Hinge	$((1 - yf)_+)^2$	✓	✓
Hinge	$(1 - yf)_+$	×	✓

As a special case, for $\widehat{\eta}_S(x)$ obtained by minimizing a suitable strongly proper loss, we have the following:

Theorem 5 (Consistency of Algorithm 1 with certain strongly proper losses). *Let D be a probability distribution on $\mathcal{X} \times \{\pm 1\}$ satisfying Assumption A. Let $\ell : \{\pm 1\} \times \bar{\mathbb{R}} \to \bar{\mathbb{R}}_+$ be a strongly proper composite loss, and let f_S, h_S denote the real-valued function and classifier learned by Algorithm 1 from a training sample S using this loss. If the kernel K and regularization parameter sequence λ_n can be chosen such that* $\text{regret}_D^{\ell}[f_S] \xrightarrow{P} 0$, *then*

$$\text{regret}_D^{\text{AM}}[h_S] \xrightarrow{P} 0 .$$

The proof of Theorem 5 involves showing that under the conditions of the theorem, the class probability estimator $\widehat{\eta}_S(x)$ in Algorithm 1 satisfies the conditions of Theorem 3 with $r = 2$; this follows as a direct consequence of the definition of strongly proper losses. Details can be found in the supplementary material.

Table 2 shows several examples of strongly proper composite losses; see (Agarwal, 2013) for more details. For each of these losses, (Zhang, 2004) gives prescriptions for K and λ_n satisfying the conditions of Theorem 5; with these choices, Algorithm 1 is AM-consistent.

5. Consistency of Empirically Balanced ERM Algorithms

Given the result of Proposition 1, another approach to optimize the AM measure is to minimize a balanced surrogate of the 0-1 loss; however this requires knowledge of p. Again, a natural approach is to use an empirical estimate $\widehat{p}_S$. This leads to a second family of algorithms that involves minimizing an *empirically* balanced loss; see Algorithm 2 for a prototype. The following establishes conditions under which such an algorithm is AM-consistent:

Theorem 6 (Consistency of Algorithm 2 with certain convex classification-calibrated losses). *Let D be a probability distribution on $\mathcal{X} \times \{\pm 1\}$ satisfying Assumption A. Let $\ell : \{\pm 1\} \times \bar{\mathbb{R}} \to \bar{\mathbb{R}}_+$ be a loss that is convex in its second argument, classification-calibrated at $\frac{1}{2}$, and for which $\exists \alpha > 0, r \geq 1$ such that $\forall \eta \in [0, 1]$, $L_\ell(\eta, 0) - H_\ell(\eta) \geq \alpha |\eta - \frac{1}{2}|^r$, and moreover $L_\ell(\frac{1}{2}, 0) = H_\ell(\frac{1}{2})$. Let f_S, h_S denote the real-valued function and classifier learned by Algorithm 2 from a training sample S using this loss. If the kernel K and regularization parameter sequence λ_n can be chosen such that* $\text{regret}_D^{\ell, (\widehat{p}_S)}[f_S] \xrightarrow{P} 0$, *then*

$$\text{regret}_D^{\text{AM}}[h_S] \xrightarrow{P} 0 .$$

Algorithm 2 Empirically Balanced ERM

1: **Input:** $S = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \{\pm 1\})^n$
2: **Select:** (a) Loss $\ell : \{\pm 1\} \times \bar{\mathbb{R}} \to \bar{\mathbb{R}}_+$; (b) RKHS $\mathcal{F}_K$ with positive definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$; (c) regularization parameter $\lambda_n > 0$
3: $\widehat{p}_S =$ (as in Eq. (2))
4: $f_S \in \text{argmin}_{f \in \mathcal{F}_K} \left\{ \frac{1}{2\widehat{p}_S n} \sum_{i: y_i = 1} \ell(1, f(x_i)) + \frac{1}{2(1 - \widehat{p}_S) n} \sum_{j: y_j = -1} \ell(-1, f(x_j)) + \lambda_n \|f\|_K^2 \right\}$
5: **Output:** Classifier $h_S(x) = \text{sign}(f_S(x))$

The proof makes use of the following lemma:[4]

Lemma 7. *Let $\ell : \{\pm 1\} \times \bar{\mathbb{R}} \to \bar{\mathbb{R}}_+$ be convex in its second argument and classification-calibrated at $\frac{1}{2}$, and suppose $\exists \alpha > 0, r \geq 1$ s.t. $\forall \eta \in [0, 1]$, $L_\ell(\eta, 0) - H_\ell(\eta) \geq \alpha |\eta - \frac{1}{2}|^r$, and $L_\ell(\frac{1}{2}, 0) = H_\ell(\frac{1}{2})$. Let $c \in (0, 1)$. For any $f : \mathcal{X} \to \bar{\mathbb{R}}$ and $h(x) = \text{sign}(f(x))$,*

$$\text{regret}_D^{0\text{-}1, (c)}[h] \leq \frac{2}{\alpha^{1/r}} \left(\text{regret}_D^{\ell, (c)}[f] \right)^{1/r} .$$

The proof of Lemma 7 makes use of some results of (Scott, 2012); details can be found in the supplementary material. The proof of Theorem 6 then follows by an application of Lemma 7 with $c = \widehat{p}_S$, which gives $\text{regret}_D^{0\text{-}1, (\widehat{p}_S)}[h_S] \xrightarrow{\text{P}} 0$; by Lemma 2, this then implies the result. Details are in the supplementary material.

Table 2 gives examples of convex classification-calibrated losses for which $\exists \alpha > 0, r \geq 1$ such that $\forall \eta \in [0, 1]$, $L_\ell(\eta, 0) - H_\ell(\eta) \geq \alpha |\eta - \frac{1}{2}|^r$, and moreover $L_\ell(\frac{1}{2}, 0) = H_\ell(\frac{1}{2})$; see (Zhang, 2004; Bartlett et al., 2006) for more details. For each of these losses, it is possible to show that K and λ_n can be chosen to satisfy the conditions of Theorem 6, yielding AM-consistency of Algorithm 2. The proof, which involves a detailed stability analysis extending the analyses in (Zhang, 2004; Bousquet & Elisseeff, 2002), is heavily technical and beyond the scope of the current paper; details will be provided in a longer version of the paper.

[4] We note that (Scott, 2012) gives a more general cost-sensitive regret bound than Lemma 7; however the bound there has an implicit dependence on c. In our case, we need a bound with an explicit dependence on c which when applied to $c = \widehat{p}_S$ and $h_S(x) = \text{sign}(f_S(x))$, yields that if $\text{regret}_D^{\ell, (\widehat{p}_S)}[f_S] \xrightarrow{\text{P}} 0$, then $\text{regret}_D^{0\text{-}1, (\widehat{p}_S)}[h_S] \xrightarrow{\text{P}} 0$.

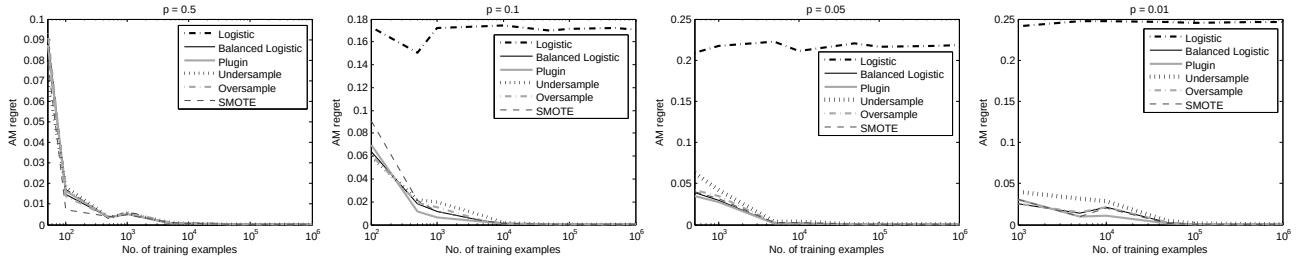

Figure 1. Experiments on synthetic data: AM-regret as a function of number of training examples, using various algorithms in conjunction with the logistic loss, for various values of class imbalance parameter p (see Section 6.1).

6. Experiments

We conducted two types of experiments to evaluate the algorithms studied in the previous sections: the first involved synthetic data from a known distribution for which the AM-regret could be calculated exactly; the second involved a large range of real data sets. We compared the algorithms with the standard underlying ERM algorithms (which seek to minimize the usual misclassification error), and also with under-sampling and over-sampling methods that seek to 'balance' imbalanced data sets and have been highly popular in class imbalance settings in practice: these include random under-sampling, in which examples from the majority class are randomly sub-sampled to equal the number of minority class examples; random over-sampling, in which examples from the minority class are randomly over-sampled to equal the number of majority class examples; and synthetic over-sampling using the SMOTE technique, in which synthetic examples from the minority class are generated along lines joining pairs of actual minority class examples in the data set (Chawla et al., 2002); in each case, the under-sampled/over-sampled data set was then given as input to the standard regularized ERM algorithm.

6.1. Synthetic Data

Our first goal was to evaluate the behavior of the AM-regret for different algorithms in a setting where it could be calculated exactly. For these experiments, we generated data in $\mathcal{X} = \mathbb{R}^d$ ($d = 10$) with varying degrees of class imbalance ($p = 0.5,\ 0.1,\ 0.05,\ 0.01$). In each case, examples in $\mathbb{R}^d \times \{\pm 1\}$ were generated as follows: each example was positive with probability p and negative with probability $1-p$, with positive instances drawn from a multivariate Gaussian distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$, and negative instances drawn from a multivariate Gaussian distribution with mean $-\mu$ and the same covariance matrix Σ; here μ was drawn uniformly at random from $\{-1, 1\}^d$, and Σ was drawn from a Wishart distribution with 20 degrees of freedom and a randomly drawn invertible PSD scale matrix. In these experiments we used the logistic loss as a prototype of a loss that satisfies the conditions of both Theorem 5 and Theorem 6. For the distributions considered, the AM-optimal classifier is linear, as are the real-valued functions minimizing the expected values of the logistic and empirically balanced logistic losses; this both makes it sufficient to learn a linear function, (i.e. to use a linear kernel), and simplifies the subsequent calculation of the AM-regret of the learned classifiers under these distributions (see supplementary material).

Figure 1 shows plots of the AM-regret as a function of the number of training examples n for different values of p for all the algorithms (all using the logistic loss; in all cases, the regularization parameter was set to $\lambda_n = 1/\sqrt{n}$). For $p = 0.5$, which corresponds to a perfectly balanced distribution, the AM-regret for all methods converges to zero. For $p < 0.5$, when the classes are imbalanced, as expected, the AM-regret of the standard logistic regression algorithm does not converge to zero; on the other hand, for both the empirical plug-in and empirically balanced algorithms, the AM-regret converges to zero. The AM-regret for the sampling methods also converges to zero; however the under-sampling method has slower convergence (since it throws away information), while the over-sampling methods are computationally expensive (since they blow up the training sample size).

6.2. Real Data

Our second goal was to evaluate the performance of the class imbalance algorithms studied here on a wide range of real data. We used 17 data sets with varying degrees of class imbalance, taken from the UCI repository (Frank & Asuncion, 2010) and other sources; due to space constraints, we discuss results on 3 of these here in detail (see Table 3; full results for all 17 data sets are included in the supplementary material).

In this case we evaluated all the above algorithms (except the random over-sampling algorithm, which for

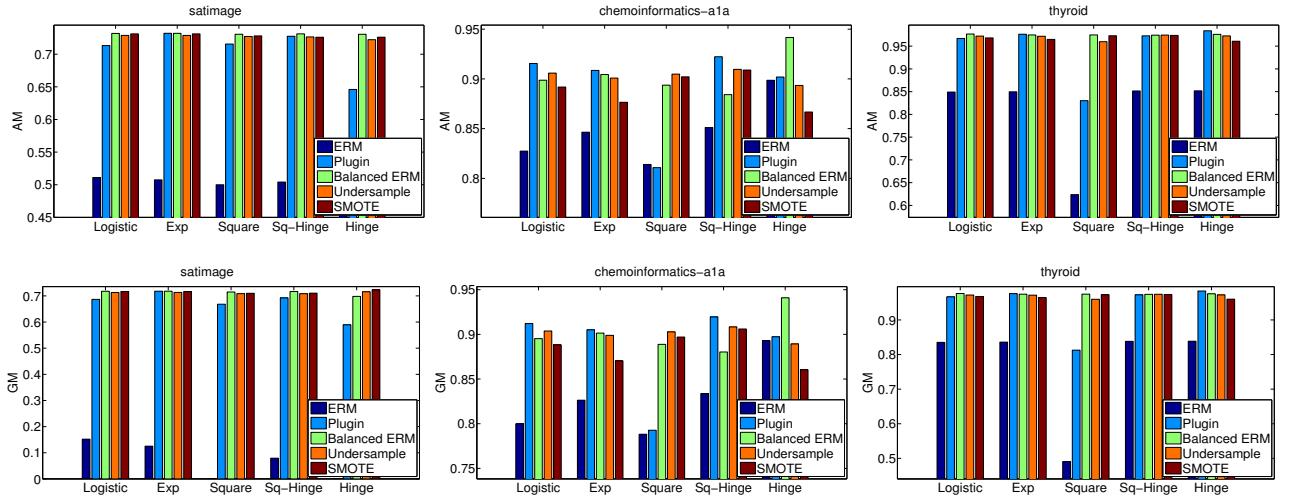

Figure 2. Results on the three data sets summarized in Table 3, using various algorithms in conjunction with different loss functions, in terms of AM (top panel) and GM (bottom panel); higher values are better (see Section 6.2).

Table 3. Summary of 3 data sets discussed here in detail; see supplementary material for details of all 17 data sets.

Data set	**# examples**	**# features**	$p = \mathbf{P}(y = 1)$
satimage	6435	36	0.0973
chemo-a1a	2142	1021	0.0233
thyroid	7200	21	0.0231

ERM-based algorithms is similar to empirical balancing but computationally more expensive), using all the five loss functions given in Table 2. In all experiments, we learned a linear function with ℓ_2 regularization; in each case, the regularization parameter λ was selected by 5-fold cross-validation on the training sample from the range $\{2^{-20}, \ldots, 2^4\}$ (the value of λ maximizing the average AM value on the validation folds was selected). For the empirical plug-in algorithm with logistic, exponential, square, and square hinge losses, which are all proper (composite) losses, class probability estimates were obtained using the standard ψ^{-1} transform based on the link function associated with the proper loss (Reid & Williamson, 2010; Agarwal, 2013); for hinge loss, we used Platt scaling (Platt, 1999).

The results, averaged over 10 random 80%-20% train-test splits for each data set (and 5 random sampling runs for the under-/over-sampling methods), are shown in Figure 2 (see supplementary material for results on additional data sets). Performance is shown in terms of AM (top panel) as well as GM (bottom panel). As expected, the standard ERM algorithms do not give good AM performance, while the empirical plug-in, empirically balanced ERM, and under-/over-sampling algorithms are mostly similar and all give good AM performance. A detailed rank analysis across the full 17 data sets suggests the empirically balanced ERM method wins overall (see supplementary material). Among loss functions, we see similar performance overall. The main exception is the square loss, which sometimes gives poor performance; this may be due to the fact that unlike other losses, it penalizes predictions with a large positive margin yf, suggesting the other losses may be preferable. Performance on the GM measure shows similar trends as for AM.

We also point out that when performance is measured in terms of AUC-ROC, we see no significant difference between the standard ERM algorithms and other methods (see supplementary material), consistent with the observation made in Table 1 that the AUC-ROC applies to a scoring function and not to a classifier, and therefore sampling and other class imbalance methods do not significantly impact the AUC-ROC.

7. Conclusion

We have studied the problem of binary classification under class imbalance, and have given the first formal consistency analysis for this problem that we are aware of. In particular, we have focused on the AM performance measure, and have shown that under certain conditions, some simple algorithms such as plug-in rules with an empirical threshold and empirically balanced ERM algorithms are AM-consistent. Our experiments confirm these findings. This suggests that at least when the AM performance measure is of interest, it may be unnecessary to throw away information as is done in under-sampling, or to incur additional computational cost as in over-sampling. A natural next step is to conduct a similar analysis for other performance measures used in class imbalance settings.

References

Agarwal, S. Surrogate regret bounds for the area under the ROC curve via strongly proper losses. In *COLT*, 2013.

Bartlett, P.L., Jordan, M.I., and McAuliffe, J.D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Bousquet, O. and Elisseeff, A. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Cardie, C. and Howe, N. Improving minority class prediction using case-specific feature weights. In *ICML*, 1997.

Chan, P.K. and Stolfo, S.J. Learning with non-uniform class and cost distributions: Effects and a distributed multi-classifier approach. In *KDD-98 Workshop on Distributed Data Mining*, 1998.

Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16: 321–357, 2002.

Chawla, N.V., Japkowicz, N., and Kolcz, A. (eds.). *Proceedings of the ICML 2003 Workshop on Learning from Imbalanced Data Sets*. 2003.

Chawla, N.V., Japkowicz, N., and Kotcz, A. Editorial: Special issue on learning from imbalanced data sets. *SIGKDD Explorations Newsletter*, 6(1):1–6, 2004.

Chawla, N.V., Cieslak, D.A., Hall, L.O., and Joshi, A. Automatically countering imbalance and its empirical relationship to cost. *Data Mining & Knowledge Discovery*, 17(2):225–252, 2008.

Cheng, J., Hatzis, C., Hayashi, H., Krogel, M-A., Morishita, S., Page, D., and Sese, J. KDD Cup 2001 report. *ACM SIGKDD Explorations Newsletter*, 3(2):47–64, 2002.

Daskalaki, S., Kopanas, I., and Avouris, N. Evaluation of classifiers for an uneven class distribution problem. *Applied Artificial Intelligence*, 20:381–417, 2006.

Davis, J. and Goadrich, M. The relationship between precision-recall and ROC curves. In *Proc. ICML*, 2006.

Drummond, C. and Holte, R.C. Severe class imbalance: Why better algorithms aren't the answer. In *Proc. ECML*, 2005.

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

Elkan, C. The foundations of cost-sensitive learning. In *Proc. IJCAI*, 2001.

Frank, A. and Asuncion, A. UCI machine learning repository, 2010. URL `http://archive.ics.uci.edu/ml`.

Gu, Q., Zhu, L., and Cai, Z. Evaluation measures of the classification performance of imbalanced data sets. In *Computational Intelligence and Intelligent Systems*, volume 51, pp. 461–471. 2009.

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

Japkowicz, N. The class imbalance problem: Significance and strategies. In *In Proc. ICAI*, 2000.

Japkowicz, N. and Stephen, S. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6 (5):429–449, 2002.

Kennedy, K., Namee, B.M., and Delany, S.J. Learning without default: a study of one-class classification and the low-default portfolio problem. In *ICAICS*, 2009.

Kotlowski, W., Dembczynski, K., and Hüllermeier, E. Bipartite ranking through minimization of univariate loss. In *Proc. ICML*, 2011.

Kubat, M. and Matwin, S. Addressing the curse of imbalanced training sets: One-sided selection. In *ICML*, 1997.

Lawrence, S., Burns, I., Back, A., Tsoi, A-C., and Giles, C.L. Neural network classification and prior class probabilities. In *Neural Networks: Tricks of the Trade*, Lecture Notes in Computer Science, pp. 1524:299–313. 1998.

Lewis, D.D. and Gale, W.A. A sequential algorithm for training text classifiers. In *Proc. SIGIR*, 1994.

Ling, C., Ling, C.X., and Li, C. Data mining for direct marketing: Problems and solutions. In *Proc. KDD*, 1998.

Liu, W. and Chawla, S. A quadratic mean based supervised learning model for managing data skewness. In *Proc. SDM*, 2011.

Platt, J.C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Adv. Large Margin Classifiers*, pp. 61–74, 1999.

Powers, R., Goldszmidt, M., and Cohen, I. Short term performance forecasting in enterprise systems. In *Proc. KDD*, 2005.

Qiao, X. and Liu, Y. Adaptive weighted learning for unbalanced multicategory classification. *Biometrics*, 65(1): 159–168, 2009.

Reid, M.D. and Williamson, R.C. Surrogate regret bounds for proper losses. In *Proc. ICML*, 2009.

Reid, M.D. and Williamson, R.C. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010.

Scott, C. Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics*, 6:958–992, 2012.

Van Hulse, J., Khoshgoftaar, T.M., and Napolitano, A. Experimental perspectives on learning from imbalanced data. In *Proc. ICML*, 2007.

Wallace, B.C., K.Small, Brodley, C.E., and Trikalinos, T.A. Class imbalance, redux. In *Proc. ICDM*, 2011.

Zadrozny, Bianca, Langford, John, and Abe, Naoki. Cost-sensitive learning by cost-proportionate example weighting. In *ICDM*, 2003.

Zhang, T. Statistical behaviour and consistency of classification methods based on convex risk minimization. *Annals of Mathematical Statistics*, 32:56–134, 2004.