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

Aditya Krishna Menon[1], Harikrishna Narasimhan[2], Shivani Agarwal[2] and Sanjay Chawla[3]

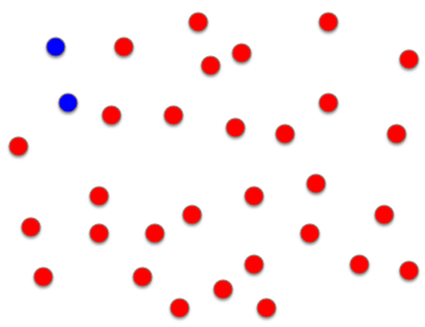
[1]University of California, San Diego [2]Indian Institute of Science, Bangalore

[3]University of Sydney and NICTA, Sydney

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, 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.

Class Imbalance



One class *rare* compared to the other!

Applications

- Medical Diagnosis
- Text Retrieval
- Credit Risk Minimization
- Fraud Detection
- Click modeling
- .....

Problem Setup

Training set: $S = ((x_1, y_1), \ldots, (x_n, y_n))$ drawn iid from D on $\mathcal{X} \times \{\pm 1\}$

Goal: Learn a binary classifier $h_S : \mathcal{X} \to \{\pm 1\}$

Imbalanced classes: $p = \mathbf{P}(y = 1)$ departs significantly from 0.5.

$$\mathrm{TPR}_D[h] = \mathbf{P}(h(x) = 1 \,|\, y = 1) \qquad \mathrm{TNR}_D[h] = \mathbf{P}(h(x) = -1 \,|\, y = -1)$$

Performance Measures

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

Two Popular Algorithms

Estimator for $p = \mathbf{P}(y = 1)$

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

Algorithm 1: Plug-in with Empirical Threshold

Learn a class probability estimator via minimization of a strongly proper loss:

$$f_S \in \mathrm{argmin}_{f \in \mathcal{F}_K} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda_n \|f\|_K^2 \right\}; \quad \widehat{\eta}_S = \psi^{-1} \circ f_S$$

Plug-in classifier: $h_S(x) = \mathrm{sign}(\widehat{\eta}_S(x) - \widehat{p}_S)$ — **Empirical Threshold**

Algorithm 2: Empirically Balanced ERM

$$f_S \in \mathrm{argmin}_{f \in \mathcal{F}_K} \left\{ \frac{1}{2\widehat{p}_S n} \sum_{i: y_i = 1} \ell(1, f(x_i)) + \frac{1}{2(1 - \widehat{p}_S) n} \sum_{j: y_j = -1} \ell(-1, f(x_j)) + \lambda_n \|f\|_K^2 \right\}$$

Empirical Balancing Terms

Classifier: $h_S(x) = \mathrm{sign}(f_S(x))$

Main Consistency Results

AM-regret: $\mathrm{regret}_D^{\mathrm{AM}}[h] = \sup_{h: \mathcal{X} \to \{\pm 1\}} \mathrm{AM}_D[h] - \mathrm{AM}_D[h]$

For any prediction space $\widehat{\mathcal{Y}} \subseteq \bar{\mathbb{R}}$, function $f : \mathcal{X} \to \widehat{\mathcal{Y}}$, and loss $\ell : \{\pm 1\} \times \widehat{\mathcal{Y}} \to \bar{\mathbb{R}}_+$,

ℓ-regret: $\mathrm{er}_D^{\ell}[f] = \mathbf{E}_{(x,y) \sim D}\left[\ell(y, f(x))\right]; \quad \mathrm{regret}_D^{\ell}[f] = \mathrm{er}_D^{\ell}[f] - \inf_{f: \mathcal{X} \to \widehat{\mathcal{Y}}} \mathrm{er}_D^{\ell}[f]$

Cost-sensitive loss: $\ell^{(c)}(y, \widehat{y}) = \left((1 - c)\,\mathbf{1}(y = 1) + c\,\mathbf{1}(y = -1)\right) \cdot \ell(y, \widehat{y})$

Theorem (Consistency of Algorithm 1 with certain strongly proper losses)

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 $\mathrm{regret}_D^{\ell}[f_S] \xrightarrow{P} 0$, then under mild conditions on the distribution D,

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

Theorem (Consistency of Algorithm 2 with certain convex classification-calibrated losses)

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 satisfies a few more technical conditions (see paper). 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 $\mathrm{regret}_D^{\ell,(\widehat{p}_S)}[f_S] \xrightarrow{P} 0$, then under mild conditions on the distribution D,

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

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

Key Ingredients in Proofs

- Balanced losses (Kotlowski et al, 2011):

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

- Decomposition lemma:

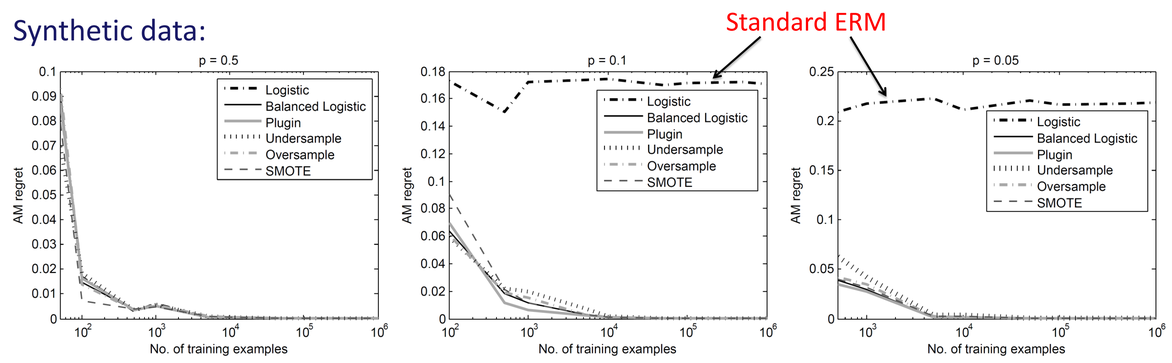
Lemma: *Let $h_S : \mathcal{X} \to \{\pm 1\}$ denote the classifier learned by an algorithm from training sample S, and let $\widehat{p}_S$ denote any estimator of $p = \mathbf{P}(y = 1)$ satisfying $\widehat{p}_S \in (0, 1)$ and $\widehat{p}_S \xrightarrow{P} p$. Then under mild conditions on the distribution D,*

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

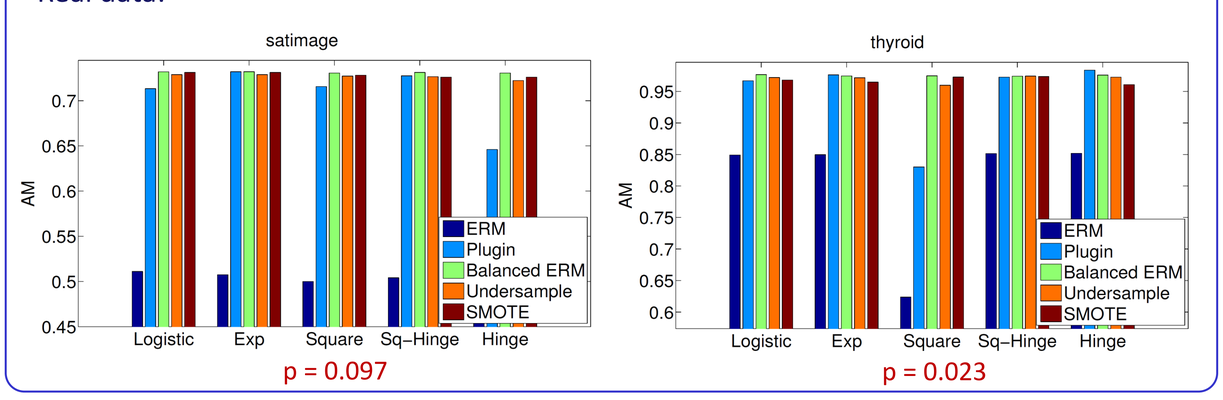
- Surrogate regret bounds for cost-sensitive classification (Scott, 2012)
- Proper and strongly proper losses (Reid and Williamson, 2009, 2010; Agarwal, 2013)
- Surrogate regret bounds for standard binary classification (Zhang, 2004; Bartlett et al, 2006)

Experiments

Synthetic data:



Real data:



References

1. S. Agarwal. Surrogate regret bounds for the area under the ROC curve via strongly proper losses. In COLT, 2013.
2. W. Kotlowski, K. Dembczynski, and E. Hüllermeier. Bipartite ranking through minimization of univariate loss. In ICML, 2011.
3. M.D. Reid and R.C. Williamson. Surrogate regret bounds for proper losses. In ICML, 2009.
4. C. Scott. Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics*, 6:958-992, 2012.