Fast imbalanced binary classification: a moment-based approach

Édouard Grave, Laurent El Ghaoui
University of California, Berkeley
{grave|elghaoui}@berkeley.edu

Abstract

In this paper, we consider the problem of imbalanced binary classification in which the number of negative examples is much larger than the number of positive examples. The two mainstream methods to deal with such problems are to assign different weights to negative and positive points or to subsample points from the negative class. In this paper, we propose a different approach: we represent the negative class by the two first moments of its probability distribution (the mean and the covariance), while still modeling the positive class by individual examples. Therefore, our formulation does not depend on the number of negative examples, making it suitable to highly imbalanced problems and scalable to large datasets. We demonstrate empirically, on a protein classification task and a text classification task, that our approach achieves similar statistical performance than the two mainstream approaches to imbalanced classification problems, while being more computationally efficient.

1. Introduction

Many real world classification problems are highly imbalanced, meaning that examples drawn from the positive class are significantly outnumbered by negative examples. Consider the problem of building a text classifier to automatically detect news article about *renewable energies*. Only a small portion of published news articles are about this topic, and the corresponding classification problem is thus highly imbalanced. Detection problems in computer vision, such as finding *cars* in images, also lead to extremely imbalanced classification problems. Indeed, the negative, or background, class is much more likely than the class of interest.

In the case of highly imbalanced datasets, a naive strategy that classifies all the examples as negative will achieves a very low classification error, since the vast

majority of examples are indeed negative. Most classifiers that minimize the classification error thus lead to decision boundaries that are skewed toward the positive class, and thus to a large false negative rate. For most imbalanced problems, this is not acceptable, since the interesting class is the positive class, and having a low false negative rate is important. Another challenge pertaining to imbalanced datasets is the fact that the number of examples from the negative class is often gigantic. This large amount of negative data points are thus the bottleneck of the optimization algorithm.

1.1. Related work

Many different approaches have been proposed to deal with imbalanced datasets, and the corresponding literature is too large to be summarized here. We invite the interested reader to look at the extensive review of the subject by He and Garcia (2009).

A first class of methods for imbalanced learning is based on *sampling*: the idea is to sample a balanced training set from the original unbalanced set of examples. Such methods are based on undersampling the negative class (Kubat et al., 1997; Batista et al., 2004), or on (synthetic) oversampling of positive examples (Chawla et al., 2002; Batista et al., 2004).

A second class of methods, refered to as cost sensitive learning, is based on assigning different misclassification costs to the negative and the positive examples (Fan et al., 1999; Drummond and Holte, 2000; Zadrozny et al., 2003).

Finally, closely related to our approach, one-class support vector machines were also considered by Raskutti and Kowalczyk (2004) in the case of extremely imbalanced datasets and for task of text classification by Manevitz and Yousef (2002).

1.2. Contributions

In this article, we propose a new formulation for solving the problem of imbalanced binary classification. Our method is inspired by the minimax approach to classification proposed by Lanckriet et al. (2003): we model the negative class by moments of its probability distribution instead of examples. Thanks to this approach our formulation does not depend on the number of negative examples, making our method suitable to highly imbalanced problems and making it scalable to large datasets. On the other hand, we propose to model the positive class by individual examples, avoiding the pitfalls mentioned above about false negative rate. More precisely, we make the following contributions:

- we propose a new formulation for the problem of imbalanced binary classification, inspired by the model of Lanckriet et al. (2003), where the negative class is represented by its mean and its variance, while the positive class is represented by individual examples (section 2);
- we give a geometric interpretation of our formulation, obtained through Lagrangian duality (section 3);
- we show that our approach is strongly related to one-class support vector machines, as introduced by Schölkopf et al. (2001) (section 4);
- we demonstrate that our approach is competitive with the two mainstream approaches to imbalanced classification problems (undersambling and asymmetric cost function) on two classification problems (sections 6 and 7).

2. Problem formulation

In this section, we propose a new formulation for solving the problem of imbalanced binary classification. We assume that the number of negatively labeled points is much larger than the number of positively labeled points. In order to cope with this large number of negative examples, we propose to model the negative class by its distribution instead of a set of examples. More precisely, we will represent the probability distribution of the negative class by its two first moments, the mean and the covariance. On the other hand, we still represent the positive class by examples.

Let $(\mathbf{x_i})_{i\in\{1,...,n\}}$ be a set of n positive training examples and let $\bar{\mathbf{x}}$ and Σ be the mean and the covariance of the probability distribution of the negative class. In the following, we will always assume that the covariance matrix Σ is positive definite. This is not a strong assumption, since we can always add a small regularization term $\lambda\mathbf{I}_d$ to the covariance matrix.

Our goal is to find the hyperplane $(\mathbf{w}, b)$ such that all the positive examples are correctly classified:

$$\mathbf{w}^\top\mathbf{x}_i - b \geq 0, \quad \forall i \in \{1, ..., n\},$$

while maximizing the probability of correctly classifying examples drawn from the negative class, with respect to all distributions with mean $\bar{\mathbf{x}}$ and covariance $\mathbf{\Sigma}$:

$$\max_{\mathbf{w},b} \inf_{\mathbf{x}\sim(\bar{\mathbf{x}},\mathbf{\Sigma})} Pr(\mathbf{w}^\top\mathbf{x} - b \leq 0), \tag{1}$$

where $\mathbf{x} \sim (\bar{\mathbf{x}}, \mathbf{\Sigma})$, refers to the class of probability distributions with mean $\bar{\mathbf{x}}$ and covariance $\mathbf{\Sigma}$. In other word, our goal is to maximizie the specificity of the separating hyperplane, while correctly classifying all the positive examples.

According to the following lemma from Lanckriet et al. (2003), which is a consequence of a theorem by Marshall et al. (1960), the condition 1 has a geometric characterization:

Lemma 1. *Let* $\bar{\mathbf{x}} \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$ *a positive definite matrix,* $\mathbf{w} \in \mathbb{R}^d$ *such that* $\mathbf{w} \neq 0$, $b \in \mathbb{R}$ *and* $\alpha \in [0, 1[$. *Then, the condition*

$$\inf_{\mathbf{x} \sim (\bar{\mathbf{x}}, \boldsymbol{\Sigma})} Pr\left(\mathbf{x}^\top \mathbf{w} - b \leq 0\right) \geq \alpha,$$

holds if and only if

$$b - \bar{\mathbf{x}}^\top \mathbf{w} \geq \kappa(\alpha) \sqrt{\mathbf{w}^\top \Sigma \mathbf{w}},$$

where $\kappa(\alpha) = \sqrt{\frac{\alpha}{1-\alpha}}$.

Using this result, the optimal hyperplane we are looking for is thus the optimal solution of the following optimization problem:

$$\max_{\alpha, \mathbf{w}, b} \quad \alpha \quad \text{s.t.} \quad b - \bar{\mathbf{x}}^\top \mathbf{w} \geq \kappa(\alpha) \sqrt{\mathbf{w}^\top \Sigma \mathbf{w}} \quad \text{and} \quad \mathbf{x}_i^\top \mathbf{w} - b \geq 0.$$

Since the function $\kappa : \alpha \mapsto \sqrt{\frac{\alpha}{1-\alpha}}$ is increasing on $[0, 1[$, this problem is equivalent to the program:

$$\max_{\kappa, \mathbf{w}, b} \quad \kappa \quad \text{s.t.} \quad b - \bar{\mathbf{x}}^\top \mathbf{w} \geq \kappa \sqrt{\mathbf{w}^\top \Sigma \mathbf{w}} \quad \text{and} \quad \mathbf{x}_i^\top \mathbf{w} - b \geq 0.$$

We can further eliminate b from this program, finally obtaining:

$$\max_{\kappa, \mathbf{w}} \quad \kappa \quad \text{s.t.} \quad \mathbf{x}_i^\top \mathbf{w} - \bar{\mathbf{x}}^\top \mathbf{w} \geq \kappa \sqrt{\mathbf{w}^\top \Sigma \mathbf{w}}. \tag{2}$$

This problem is a convex program (Boyd and Vandenberghe, 2004), which has an interesting geometric interpretation that will discuss in the next section. We will also show that this formulation is strongly related to one-class support vector machines (Schölkopf et al., 2001).

3. Dual problem and geometric interpretation

In this section, we derive the dual of the problem defined in equation 2 through Lagragian duality. This will allow us to give a nice geometric interpretation of our approach. Assuming that Σ is invertible, we can perform the change of variable $\mathbf{u} = \Sigma^{1/2} \mathbf{w}$. Then, our problem becomes:

$$\max_{\kappa, \mathbf{u}} \kappa \quad \text{s.t.} \quad \mathbf{z}_i^\top \mathbf{u} - \bar{\mathbf{z}}^\top \mathbf{u} \geq \kappa \sqrt{\mathbf{u}^\top \mathbf{u}},$$

where $\mathbf{z}_i = \Sigma^{-1/2}\mathbf{x}_i$ and $\bar{\mathbf{z}} = \Sigma^{-1/2}\bar{\mathbf{x}}$. Exploiting the homogeneity in variable $\mathbf{u}$, we can impose the constraint $\|\mathbf{u}\|_2 = 1$, leading to the program

$$\max_{\kappa, \mathbf{u}} \quad \kappa \quad \text{s.t.} \quad \mathbf{z}_i^\top \mathbf{u} - \bar{\mathbf{z}}^\top \mathbf{u} \geq \kappa \quad \text{and} \quad \|\mathbf{u}\|_2 = 1.$$

This problem is equivalent to

$$\max_{\kappa, \mathbf{u}} \quad \kappa \quad \text{s.t.} \quad \mathbf{z}_i^\top \mathbf{u} - \bar{\mathbf{z}}^\top \mathbf{u} \geq \kappa \quad \text{and} \quad \|\mathbf{u}\|_2 \leq 1.$$

Then, introducing dual variables p_i, it can be expressed as

$$\max_{\kappa, \mathbf{u}} \min_{p_i \geq 0} \quad \kappa + \sum_i p_i \left[(\mathbf{z}_i - \bar{\mathbf{z}})^\top \mathbf{u} - \kappa \right] \quad \text{s.t.} \quad \|\mathbf{u}\|_2 \leq 1.$$

Using Sion's minimax theorem (Sion et al., 1957), we can invert the min and the max operators, obtaining:

$$\min_{p_i \geq 0} \max_{\kappa, \mathbf{u}} \quad \kappa + \sum_i p_i \left[(\mathbf{z}_i - \bar{\mathbf{z}})^\top \mathbf{u} - \kappa \right] \quad \text{s.t.} \quad \|\mathbf{u}\|_2 \leq 1.$$

Then, if $\sum_i p_i \neq 1$, the max over κ is equal to $+\infty$. The previous program is thus equivalent to

$$\min_{p_i \geq 0} \max_{\mathbf{u}} \quad \sum_i p_i (\mathbf{z}_i - \bar{\mathbf{z}})^\top \mathbf{u} \quad \text{s.t.} \quad \|\mathbf{u}\|_2 \leq 1 \quad \text{and} \quad \sum_i p_i = 1.$$

Finally, using the definition of the dual norm, and the fact that the dual norm of the ℓ_2-norm is itself, we get

$$\min_{p_i \geq 0} \quad \| \sum_i p_i \mathbf{z}_i - \bar{\mathbf{z}} \|_2 \quad \text{s.t.} \quad \sum_i p_i = 1.$$

We can now replace $\mathbf{z}_i$ and $\bar{\mathbf{z}}$ by their expressions, to obtain the dual problem

$$\min_{p_i \geq 0} \quad \| \sum_i p_i \mathbf{x}_i - \bar{\mathbf{x}} \|_{\Sigma^{-1}} \quad \text{s.t.} \quad \sum_i p_i = 1,$$

where $\| \cdot \|_{\Sigma^{-1}}$ is the norm defined by:

$$\|\mathbf{x}\|_{\Sigma^{-1}} \;=\; \sqrt{\mathbf{x}^\top \Sigma^{-1} \mathbf{x}}.$$

Thus, the dual problem correspond to finding the orthogonal projection of the mean vector $\bar{\mathbf{x}}$ of the negative class onto the convex hull of the positive examples, for the

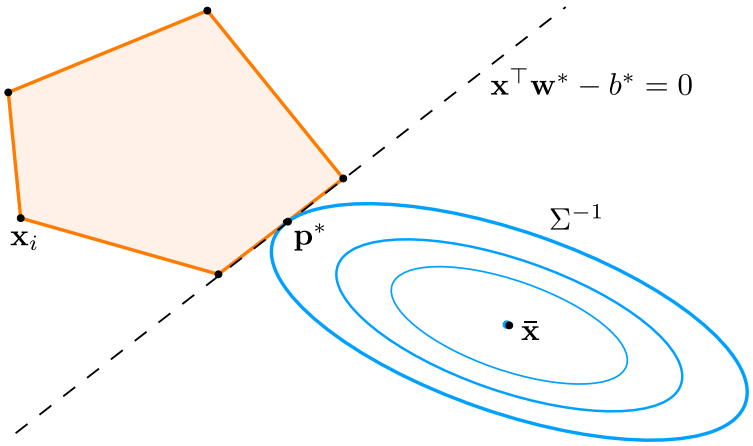

Figure 1: Geometric interpretation of our proposed formulation for imbalanced binary classification.

scalar product defined by the inverse covariance matrix Σ^{-1}. Given an optimal solution $\mathbf{p}^*$ of the dual problem, the correponding primal solution is equal to

$$\mathbf{w}^* = \Sigma^{-1}(\mathbf{X}\mathbf{p}^* - \bar{\mathbf{x}}).$$

Thus, the corresponding separating hyperplane is orthogonal to the difference between $\bar{\mathbf{x}}$ and its projection on the convex hull of the positive points $\mathbf{x}_i$, for the scalar product defined by the inverse covariance matrix Σ^{-1}. This geometric interpretation is illustrated in Figure 1.

4. Relation to one-class support vector machines

In this section, we show that our approach is strongly related to the one-class support vector machine formulation introduced by Schölkopf et al. (2001). First, let us restate our formulation for solving the problem of imbalanced binary classification:

$$\max_{\kappa, \mathbf{w}} \quad \kappa \quad \text{s.t.} \quad \mathbf{x}_i^\top \mathbf{w} - \bar{\mathbf{x}}^\top \mathbf{w} \geq \kappa \sqrt{\mathbf{w}^\top \Sigma \mathbf{w}}.$$

Exploiting the homogeneity in $\mathbf{w}$ again, we now impose that $\kappa\sqrt{\mathbf{w}^\top \Sigma \mathbf{w}} = 1$, leading to the formulation

$$\max_{\mathbf{w}, \kappa} \quad \kappa \quad \text{s.t.} \quad (\mathbf{x}_i - \bar{\mathbf{x}})^\top \mathbf{w} \geq 1 \quad \text{and} \quad \kappa = \frac{1}{\sqrt{\mathbf{w}^\top \Sigma \mathbf{w}}}.$$

We can now eliminate the variable κ, and since the function $x \mapsto 1/\sqrt{x}$ is decreasing on $\mathbb{R}_+^*$, we obtain the equivalent program:

$$\min_{\mathbf{w}} \quad \mathbf{w}^\top \Sigma \mathbf{w} \quad \text{s.t.} \quad (\mathbf{x}_i - \bar{\mathbf{x}})^\top \mathbf{w} \geq 1.$$

This formulation is very close to the formulation of hard margin, one-class support vector machines, with two important differences: first, we do not minimize the ℓ_2-norm of $\mathbf{w}$, but the Mahalanobis norm corresponding to the covariance matrix of the negative class distribution. Second, we do not try to separate the positively labeled points from the origin, but from the mean of the negative class distribution.

Similarly to support vector machines, the constraint that all positive examples should be correctly classified might be unrealistic in practice. We thus propose to relax these constraints by introducing the penalized slack variables $\xi_i \geq 0$. We then obtain the following convex program:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}\mathbf{w}^\top \Sigma \mathbf{w} + C \sum_{i=1}^{n} \xi_i \\ \text{s.t} \quad & (\mathbf{x}_i - \bar{\mathbf{x}})^\top \mathbf{w} \geq 1 - \xi_i, \quad \xi_i \geq 0, \end{aligned} \tag{3}$$

where $C \in \mathbb{R}_+$ is a tradeoff parameter. This formulation is the one that we will use in practice, and will be refered to as moment-based imbalanced binary classifier, or MIBC.

Again, this formulation is strongly related to the formulation of soft margin, one-class support vector machines. Indeed, by making the change of variables $\mathbf{u} = \Sigma^{1/2}\mathbf{w}$, and by introducing vectors $\mathbf{z}_i$ such that $\mathbf{z}_i = \Sigma^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}})$, we obtain the formulation

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}\mathbf{u}^\top \mathbf{u} + C \sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & \mathbf{z}_i^\top \mathbf{u} \geq 1 - \xi_i, \quad \xi_i \geq 0 \end{aligned}$$

which is exactly the formulation of one-class support vector machines.

5. Solving the optimization problem

In this section, we discuss how to optimize our proposed formulation for imbalanced binary classification. As said earlier, one major difference between our approach and (one-class) support vector machines is the fact that we minimize the Mahalanobis norm corresponding to the covariance matrix Σ of the negative class

distribution. Thus, the dimension of the problem will be an important factor in the choice of the optimization algorithm for our method. We will thus discuss optimization algorithms for small dimensional problems first (problems where the dimension is smaller than the number of positive data points), before moving to high dimensional problems. But first, let us derive the dual of the problem introduced in equation 3, as it will be usefull in our discussion.

5.1. Dual problem

One of the classical ways to compute the solution of a support vector machine is to solve the dual problem (Platt et al., 1998). We will follow a similar approach. Thus, we now compute the dual problem of our our moment-based imbalanced binary classifier, as defined in equation 3. Let us define the vectors $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}}$. Introducing Lagrange multipliers $\alpha \in \mathbb{R}^n$ and $\nu \in \mathbb{R}^n$, such that $\alpha_i \geq 0$ and $\nu_i \geq 0$, the Lagrangian corresponding to the problem 3 is equal to

$$\mathcal{L}(\mathbf{w}, \xi, \alpha, \nu) = \frac{1}{2}\mathbf{w}^\top \Sigma \mathbf{w} + C\sum_{i=1}^{n} \xi_i - \sum_i \alpha_i \left[\tilde{\mathbf{x}}_i^\top \mathbf{w} - 1 + \xi_i\right] - \nu^\top \xi.$$

To find the dual function, we minimize the Lagrangian over $\mathbf{w}$ and ξ. Minimizing over ξ, we find that the dual function is equal to $-\infty$, unless $C - \alpha_i - \nu_i = 0$, in which case, we are left with

$$\mathcal{L}(\mathbf{w}, \xi, \alpha, \nu) = \frac{1}{2}\mathbf{w}^\top \Sigma \mathbf{w} - \sum_i \alpha_i \left[\tilde{\mathbf{x}}_i^\top \mathbf{w} - 1\right].$$

Minimizing over $\mathbf{w}$, we then obtain that

$$\mathbf{w} = \Sigma^{-1}\tilde{\mathbf{X}}\alpha,$$

where $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, ..., \tilde{\mathbf{x}}_n] = [\mathbf{x}_1 - \bar{\mathbf{x}}, ..., \mathbf{x}_n - \bar{\mathbf{x}}]$. Replacing $\mathbf{w}$ by its optimal value and combining the constraints $\nu_i \geq 0$ with $C - \alpha_i - \nu_i = 0$, we finally obtain the dual problem

$$\max_{0 \leq \alpha \leq C} -\frac{1}{2}\alpha^\top \tilde{\mathbf{X}}^\top \Sigma^{-1} \tilde{\mathbf{X}}\alpha + \mathbf{1}^\top \alpha. \tag{4}$$

Unsurprisingly, this dual is very similar to the dual of support vector machines. Thus, most algorithms that were proposed to solve the dual of SVMs could be used to solve this dual.

5.2. Small dimensional problems

Similarly to support vector machines, the dual problem obtained in equation 4 is a quadratic program of dimension n, where n is the number of positive points. We

first need to compute the inverse of the covariance matrix Σ. This operation has a complexity of $O(d^3)$, where d is the dimension of the data. Then, finding the optimal solution of the dual problem has a complexity of $O(n^3)$. Thus, the overall complexity is $O(d^3 + n^3)$, and this approach is only applicable in the case of small dimensional problem. In the following, we will assume that the covariance matrix Σ has some structure that can be exploited to solve high dimensional problems.

5.3. High dimensional problems: factor models

In this section, we describe how to speed up the computation in the case of a factor model, that is when $\Sigma = \mathbf{D} + \mathbf{F}\mathbf{F}^\top$, where $\mathbf{D}$ is a positive definite diagonal matrix and $\mathbf{F} \in \mathbb{R}^{d \times k}$ with $k \ll d$. We will assume, without loss of generality, that $\mathbf{D} = \lambda \mathbf{I}_d$. Replacing Σ by its new expression, the dual problem obtained in equation 4 then become:

$$\max_{0 \leq \alpha \leq C} -\frac{1}{2}\alpha^\top \tilde{\mathbf{X}}^\top (\lambda \mathbf{I}_d + \mathbf{F}\mathbf{F}^\top)^{-1} \tilde{\mathbf{X}}\alpha + \alpha^\top \mathbf{1}.$$

Using the Woodbury matrix identity, we then obtain the equivalent problem

$$\max_{0 \leq \alpha \leq C} -\frac{1}{2\lambda}\alpha^\top \tilde{\mathbf{X}}^\top \left(\mathbf{I}_d - \mathbf{F}(\lambda \mathbf{I}_k + \mathbf{F}^\top \mathbf{F})^{-1}\mathbf{F}^\top\right) \tilde{\mathbf{X}}\alpha + \alpha^\top \mathbf{1}.$$

We thus have replaced the inverse of a $d \times d$ matrix by the inverse of a $k \times k$ matrix. The most expensive step to compute the matrix appearing in the dual cost function is to compute the matrix product $\mathbf{F}^\top \tilde{\mathbf{X}}$. The complexity of this operation is $O(kdn)$. Then, solving the dual has a complexity of $O(n^3)$.

Another potential approach to speed up the optimization algorithm for high dimensional settings is to directly estimate a sparse inverse covariance matrix (Friedman et al., 2008; d'Aspremont et al., 2008). Recently, Rolfs et al. (2012) and Hsieh et al. (2013) have proposed efficient algorithms to solve this problem, making it possible to estimate sparse inverse covariance matrices in high dimensional spaces.

Finally, it should be noted that it is often the case that several problems with the same covariance matrix have to be solved, for example when doing a grid search for parameter selection. Thus, the computation or the estimation of the inverse of the covariance matrix (or the computation of a factor model) is often amortized over the resolution of several problems.

6. Small scale experiments

In this section, we present experiments performed on small dimensional datasets. We compare our moment-based imbalanced binary classifier (MIBC) with two

standard strategies for imbalanced classification problems, using SVMs: undersambling the negative class and using different costs for the negative and positive examples. We will use the liblinear (Fan et al., 2008) implementation of linear support vector machines, and an implementation of our approach using Mosek[1]. It is thus important to note that our implementation is based on a general purpose quadratic programming solver. A custom implementation, for example based on the SMO algorithm, should greatly improve the efficiency of our moment-based imbalanced binary classifier.

Dataset	# positive	# negative	ratio
PHOSS	613	10,798	17
PHOST	140	9,051	64
PHOSY	136	5,103	37
CAM	942	17,974	19

Table 1: Basic statistics about the different datasets.

6.1. Datasets

We evaluated the different methods on four datasets introduced by Radivojac et al. (2004), which are publicly available[2]. These datasets correspond to protein classification problems, such as predicting protein phosphorylation sites (PHOST, PHOSS, PHOSY) or predicting binding regions (CAM). Following the approach proposed by Radivojac et al. (2004), we keep the 150 features which are the most correlated to the class labels. The ratio of negative examples to positive ones varies from 17.6 to 64.6 on the different datasets. Basic statistics about those are given in Table 1.

6.2. Methodology

For each dataset, we use 50% of the examples as training set, 20% as validation set and 30% as test set. For all methods, we chose C in the set $\{10^5, ..., 10^{-4}\}$. When undersampling the negative class, we keep as many negative examples as positive examples. For the asymmetric cost function method, we consider the following ratios between the weights for the positive examples and negative examples: $\{1.0, r/4, r/2, r, 2r\}$, where r is the ratio of number of negative examples to the number positive ones. We replicated the experiments over twenty random splits of the data.

[1] www.mosek.com

[2] www.informatics.indiana.edu/predrag/publications.htm

	This work	Cost-sensitive	Sampling
PHOSS	$77.2^{\dagger} \pm 0.7$	76.8 ± 0.8	74.3 ± 1.1
PHOST	$77.4^{\dagger} \pm 1.7$	73.0 ± 2.0	72.0 ± 1.5
PHOSY	$76.2^{\dagger} \pm 1.5$	72.8 ± 1.7	70.1 ± 2.1
CAM	78.2 ± 0.5	78.1 ± 0.5	75.3 ± 0.4

Table 2: Areas under the ROC curve (with confidence intervals), averaged over twenty experiments, on the protein classification tasks. † indicates that our method is significantly better than the two others (with p-value $p < 0.01$).

	This work	Cost-sensitive	Speed-up
PHOSS	146	325	$2.2\times$
PHOST	23	112	$4.8\times$
PHOSY	19	41	$2.1\times$
CAM	425	605	$1.4\times$

Table 3: Computational times, in milliseconds, required to solve one protein classification problem, averaged over twenty experiments.

6.3. Discussion

We report areas under the ROC curve for the four datasets in Table 2, computational times in Table 3 and ROC curves for two datasets in Figure 2 (PHOST and PHOSY). We performed a paired samples t-test to determine if our results are statistically significant.

First, we observe that our moment-based imbalanced binary classifier always outperforms the undersampling approach, while performing at least as well as the cost sensitive method. Second, the two datasets on which our method outperforms the asymmetric cost function SVM (PHOST and PHOSY) correspond to the highest ratio of number of negative to positive examples (64 and 37 respectively). This seems to indicate that our method is particularly adapted to highly imbalanced datasets. Finally, our method is computationally more efficient, leading to speed-up between 1.4 and 4.8 over cost-sensitive SVM, while obtaining as good or even better statistical performances. We remind our reader that we implemented our method using Python and Mosek, and it is thus certainly possible to get much better performances.

Topic	This work	Cost-sensitive	Sampling
2	89.7 ± 1.0	89.9 ± 1.4	87.7 ± 1.2
9	96.1 ± 0.7	96.3 ± 0.8	94.1 ± 1.3
25	95.1 ± 0.8	94.3 ± 1.6	93.7 ± 1.2
33	96.0 ± 0.4	95.7 ± 0.6	93.9 ± 0.7
59	96.1 ± 0.4	95.9 ± 1.4	95.0 ± 0.6
84	96.9 ± 0.8	96.4 ± 1.5	96.3 ± 0.9

Table 4: Areas under the ROC curve (with confidence intervals), averaged over ten experiments, on the text classification tasks. Differences between our moment-based imbalanced binary classifier and the subsampling method are statistically significant (with p-value $p < 0.01$).

7. Application to text classification

In this section, we report experiments performed on the task of text classification. We will follow the same methodology as described in the section 6.2. Since bag-of-words representations of textual documents live in high dimensional spaces, we propose to replace the full covariance matrix of the negative class by its diagonal.

7.1. Dataset

We use the REUTERS RCV1 dataset, introduced by Lewis et al. (2004), which is a classical test bed for text classification methods. Each document of the corpus is tagged with respect to three different category sets: topics, industries and regions. We consider classification problems that consist in classifying documents that are labeled with a given topic label *v.s.* the rest of the documents. There are 104 different topics, and we will thus consider only a subset of the 104 possible classification tasks. Since we want to focus on highly imbalanced classification problems, we set the ratio of negative examples to positive examples to $1,000$.

7.2. Discussion

We report areas under the ROC curve in Table 4, computational times in Table 5 and ROC curves in Figure 3. We performed a paired samples t-test to determine if our results are statistically significant.

We observe that our moment-based imbalanced binary classifier achieves similar statistical performances than the cost-sensitive method, while generally outperforming the undersampling approach. In particular, we observe that on the classification problem TOPIC 25 V.S. REST our method achieves a much lower false

Topic	This work	Cost-sensitive	Speed-up
2	33	1088	33×
9	49	1451	29×
25	56	1211	21×
33	74	1788	24×
59	62	1299	21×
84	56	2056	36×

Table 5: Computational times, in milliseconds, required to solve one text classification problem, averaged over ten experiments.

positive rate, for a true positive rate equal to 1.0 (See Figure 3). Finally, our approach to imbalanced classification is much more computationally efficient than a SVM with asymmetric costs, leading to speed-up between 21 and 36.

8. Conclusion

In this paper, we introduced a new approach to imbalanced classification problems, refered to as moment-based imbalanced binary classifier (MIBC). We proposed to cope with the large number of negative examples by modeling the negative class by the mean and the variance of the corresponding probability distribution. On the other hand, we still model the positive class by individual examples. We show that our formulation is strongly related to one-class support vector machines. We empirically demonstrated that our approach is competitive with other methods for imbalanced classification, based on subsampling and cost sensitive learning. In particular, we demonstrated that our approach leads to much lower computational times than cost-sensitive learning, while obtaining as good, or even better, classification performances. On the other hand, our method achieves better statistical performances than subsampling.

As future work, we would like to apply our technique to detection problems in computer vision, such as finding cars or pedestrians in images. We would also like to implement efficient optimization algorithms for our approach, in order to fully exploit the much reduced computational complexity of our approach, compared to cost-sensitive support vector machines.

Acknowledgments

Édouard Grave is supported by a grant from Inria (Associated-team STATWEB).

References

Batista, G. E., Prati, R. C., and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter*, 6(1):20–29.

Boyd, S. P. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:341–378.

d'Aspremont, A., Banerjee, O., and El Ghaoui, L. (2008). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30(1):56–66.

Drummond, C. and Holte, R. C. (2000). Exploiting the cost (in) sensitivity of decision tree splitting criteria. In *ICML*, pages 239–246.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

Fan, W., Stolfo, S. J., Zhang, J., and Chan, P. K. (1999). Adacost: misclassification cost-sensitive boosting. In *ICML*, pages 97–105. Citeseer.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

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

Hsieh, C.-J., Sustik, M. A., Dhillon, I., Ravikumar, P., and Poldrack, R. (2013). Big & quic: Sparse inverse covariance estimation for a million variables. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *NIPS*.

Kubat, M., Matwin, S., et al. (1997). Addressing the curse of imbalanced training sets: one-sided selection. In *ICML*, volume 97, pages 179–186.

Lanckriet, G. R., Ghaoui, L. E., Bhattacharyya, C., and Jordan, M. I. (2003). A robust minimax approach to classification. *The Journal of Machine Learning Research*, 3:555–582.

Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397.

Manevitz, L. M. and Yousef, M. (2002). One-class svms for document classification. *the Journal of machine Learning research*, 2:139–154.

Marshall, A. W., Olkin, I., et al. (1960). Multivariate chebyshev inequalities. *The Annals of Mathematical Statistics*, 31(4):1001–1014.

Platt, J. et al. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines.

Radivojac, P., Chawla, N. V., Dunker, A. K., and Obradovic, Z. (2004). Classification and knowledge discovery in protein databases. *Journal of Biomedical Informatics*, 37(4):224–239.

Raskutti, B. and Kowalczyk, A. (2004). Extreme re-balancing for svms: a case study. *ACM Sigkdd Explorations Newsletter*, 6(1):60–69.

Rolfs, B., Rajaratnam, B., Guillot, D., Wong, I., and Maleki, A. (2012). Iterative thresholding algorithm for sparse inverse covariance estimation. In *NIPS*, pages 1574–1582.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471.

Sion, M. et al. (1957). *General Minimax Theorems*. United States Air Force, Office of Scientific Research.

Zadrozny, B., Langford, J., and Abe, N. (2003). Cost-sensitive learning by cost-proportionate example weighting. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 435–442. IEEE.

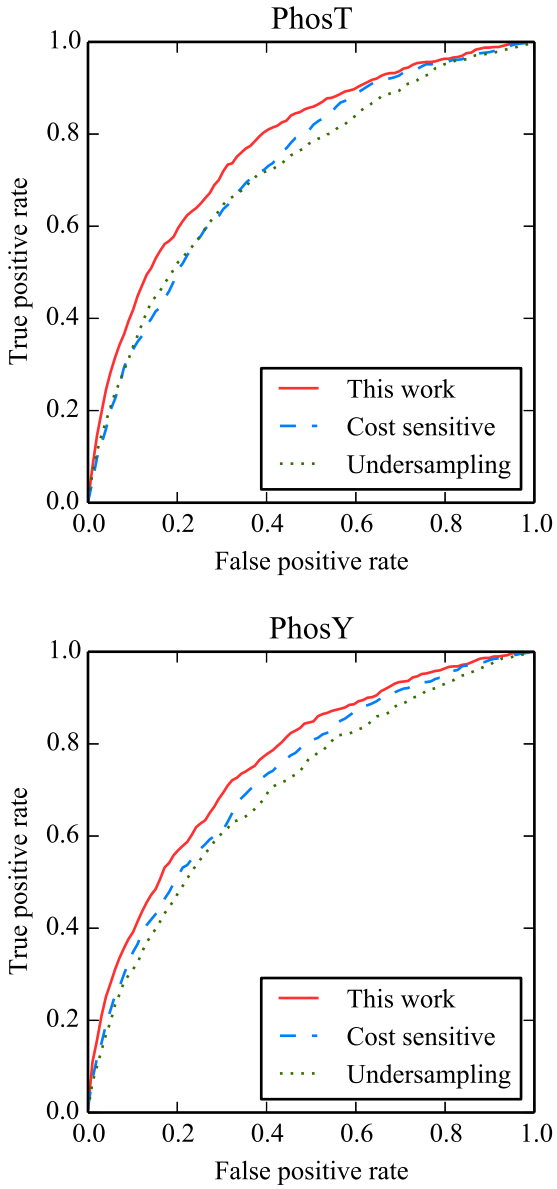

Figure 2: ROC curves, averaged over twenty experiments, on the PHOST and PHOSY datasets.

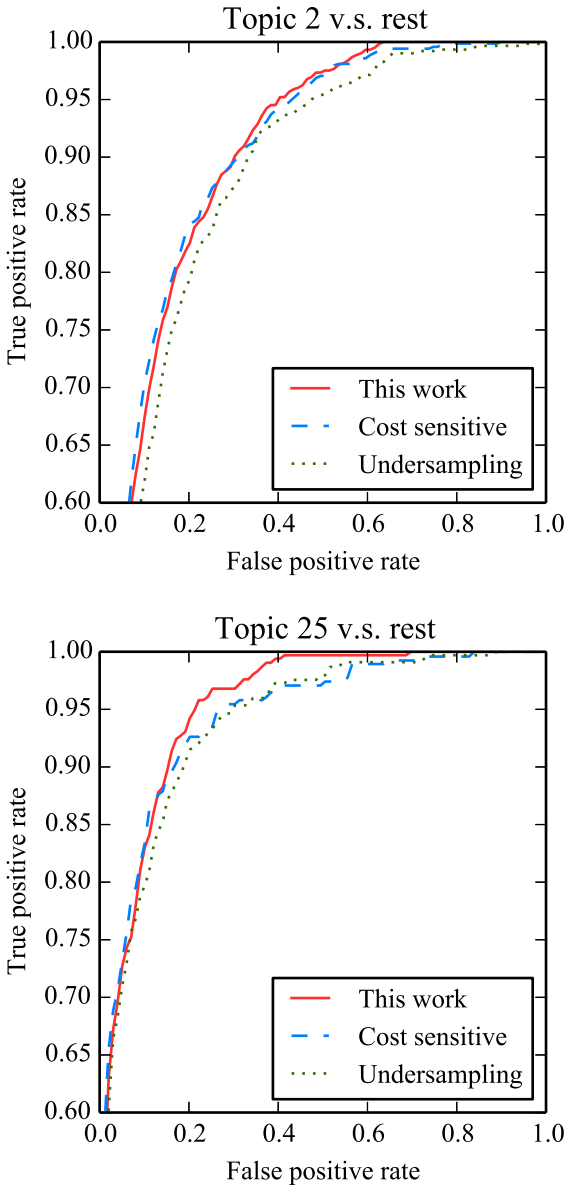

Figure 3: ROC curves, averaged over ten experiments, on the REUTERS RCV1 dataset.