Risk minimization, probability elicitation, and cost-sensitive SVMs

Hamed Masnadi-Shirazi HMASNADI@UCSD.EDU
Nuno Vasconcelos NUNO@UCSD.EDU
Statistical Visual Computing Laboratory, University of California, San Diego, La Jolla, CA 92039

Abstract

A new procedure for learning cost-sensitive SVM classifiers is proposed. The SVM hinge loss is extended to the cost sensitive setting, and the cost-sensitive SVM is derived as the minimizer of the associated risk. The extension of the hinge loss draws on recent connections between risk minimization and probability elicitation. These connections are generalized to cost-sensitive classification, in a manner that guarantees consistency with the cost-sensitive Bayes risk, and associated Bayes decision rule. This ensures that optimal decision rules, under the new hinge loss, implement the Bayes-optimal cost-sensitive classification boundary. Minimization of the new hinge loss is shown to be a generalization of the classic SVM optimization problem, and can be solved by identical procedures. The resulting algorithm avoids the shortcomings of previous approaches to cost-sensitive SVM design, and has superior experimental performance.

1. Introduction

The most popular strategy for the design of classification algorithms is to minimize the probability of error, assuming that all misclassifications have the same cost. The resulting decision rules are usually denoted as *cost-insensitive*. However, in many important applications of machine learning, such as medical diagnosis, fraud detection, or business decision making, certain types of error are much more costly than others. Other applications involve significantly unbalanced datasets, where examples from different classes appear with substantially different probability. It is well known, from Bayesian decision theory, that under any of these two situations (uneven costs or probabilities), the optimal decision rule deviates from the optimal cost-insensitive rule in the same manner. In both cases, reliance on cost insensitive algorithms for classifier design can be highly sub-optimal. While this makes it obviously important to develop *cost-sensitive* extensions of state-of-the-art machine learning techniques, the current understanding of such extensions is limited.

In this work we consider the support vector machine (SVM) architecture (Cortes & Vapnik, 1995). Although SVMs are based on a very solid learning-theoretic foundation, and have been successfully applied to many classification problems, it is not well understood how to design cost-sensitive extensions of the SVM learning algorithm. The standard, or cost-insensitive, SVM is based on the minimization of a symmetric loss function (the hinge loss) that does not have an obvious cost-sensitive generalization. In the literature, this problem has been addressed by various approaches, which can be grouped into three general categories. The first is to address the problem as one of data processing, by adopting resampling techniques that under-sample the majority class and/or over-sample the minority class (Kubat & Matwin, 1997; Chawla et al., 2002; Akbani et al., 2004). Resampling is not easy when the classification unbalance is due to either different misclassification costs (not clear what the class probabilities should be) or an extreme unbalance in class probabilities (sample starvation for classes of very low probability). It also does not guarantee that the learned SVM will change, since it could have no effect on the support vectors. The second class of approaches (Amari & Wu, 1999; Wu & Chang, 2003; 2005) involves kernel modifications. These methods are based on conformal transformations of the input or feature space, by modifying the kernel used by the SVM. They are somewhat unsatisfactory, due to the implicit assumption that a linear SVM cannot be made cost-sensitive. It is unclear why this should be the case.

The third, and most widely researched, approach is to modify the SVM algorithm in order to achieve cost sensitivity. This is done in one of two ways. The first is a naive method, known as *boundary movement (BM-SVM),* which shifts the decision boundary by simply adjusting the threshold of the standard SVM (Karakoulas & Shawe-Taylor, 1999). Under Bayesian decision theory, this would be the optimal

Appearing in *Proceedings of the 27^{th} International Conference on Machine Learning*, Haifa, Israel, 2010. Copyright 2010 by the author(s)/owner(s).

strategy if the class posterior probabilities were available. However, it is well known that SVMs do not predict these probabilities accurately. While a literature has developed in the area of probability calibration (Platt, 2000; Elkan, 2001), calibration techniques do not aid the cost-sensitive performance of threshold manipulation. This follows from the fact that all calibration techniques rely on an invertible (monotonic and one-to-one) transformation of the SVM output. Because the manipulation of a threshold at either the input or output of such a transformation produces the same receiver-operating-characteristic (ROC) curve, calibration does not change cost-sensitive classification performance. The boundary movement method is also obviously flawed when the data is non-separable, in which case cost-sensitive optimality is expected to require a modification of *both* the normal of the separating plane w and the classifier threshold b. The second proposal to modify SVM learning is known as the *biased penalties (BP-SVM)* method (Bach et al., 2006; Lin et al., 2002; Davenport et al., 2006; Wu & Srihari, 2003; Chang & Lin, 2001). This consists of introducing different penalty factors C_1 and C_{-1} for the positive and negative SVM slack variables during training. It is implemented by transforming the primal SVM problem into

$$\arg\min_{w,b,\xi} \frac{1}{2}||w||^2 + C\left[C_1 \sum_{\{i|y_i=1\}} \xi_i + C_{-1} \sum_{\{i|y_i=-1\}} \xi_i\right]$$
$$\text{s.t. } y_i(w^T x + b) \geq 1 - \xi_i. \tag{1}$$

The biased penalties method also suffers from an obvious flaw, which is converse to that of the boundary movement method: it has limited ability to enforce cost-sensitivity when the training data is separable. For large slack penalty C, the slack variables ξ_i are zero-valued and the optimization above degenerates into that of the standard SVM, where the decision boundary is placed midway between the two classes (rather than assigning a larger margin to one of them).

In this work we propose an alternative strategy for the design of cost-sensitive SVMs. This strategy is fundamentally different from previous attempts, in the sense that is does not directly manipulate the standard SVM learning algorithm. Instead, we extend the SVM hinge loss, and derive the optimal cost-sensitive learning algorithm as the minimizer of the associated risk. The derivation of the new cost-sensitive hinge loss draws on recent connections between risk minimization and probability elicitation (Masnadi-Shirazi & Vasconcelos, 2009). Such connections are generalized to the case of cost-sensitive classification.

It is shown that it is always possible to specify the predictor and conditional risk functions desired for the SVM classifier, and derive the loss for which these are optimal. A sufficient condition for the cost-sensitive Bayes-optimality of the predictor is then provided, as well as necessary conditions for conditional risks that approximate the cost-sensitive Bayes risk. Together, these conditions enable the design of a new hinge loss which is minimized by an SVM that 1) implements the cost-sensitive Bayes decision rule, and 2) approximates the cost-sensitive Bayes risk. It is also shown that the minimization of this loss is a generalization of the classic SVM optimization problem, and can be solved by identical procedures. The resulting algorithm avoids the shortcomings of previous methods, producing cost-sensitive decision rules for *both* cases of separable and inseparable training data. Experimental results show that these advantages result in better cost-sensitive classification performance than previous solutions.

The paper is organized as follows. Section 2 briefly reviews the probability elicitation view of loss function design (Masnadi-Shirazi & Vasconcelos, 2009). Section 3 then generalizes the connections between probability elicitation and risk minimization to the cost-sensitive setting. In Section 4, these connections are used to derive the new SVM loss and algorithm. Finally, Section 5 presents an experimental evaluation that demonstrates improved performance of the proposed cost sensitive SVM over previous methods.

2. Probability elicitation and the risk

A classifier h maps a feature vector $\mathbf{x} \in \mathcal{X}$ to a class label $y \in \{-1, 1\}$. This mapping can be written as $h(\mathbf{x}) = sign[f(\mathbf{x})]$ for some function $f : \mathcal{X} \to \mathbb{R}$, which is denoted as the classifier predictor. Feature vectors and class labels are drawn from probability distributions $P_{\mathbf{X}}(\mathbf{x})$ and $P_Y(y)$ respectively. Given a non-negative loss function $L(\mathbf{x}, y)$, the classifier is optimal if it minimizes the risk $R(f) = E_{\mathbf{X},Y}[L(h(\mathbf{x}), y)]$. This is equivalent to minimizing the conditional risk

$$\begin{aligned} E_{Y|\mathbf{X}}[L(h(\mathbf{x}), y)|\mathbf{X} = \mathbf{x}] = P_{Y|\mathbf{X}}(1|\mathbf{x})L(h, 1) \\ +(1 - P_{Y|\mathbf{X}}(1|\mathbf{x}))L(h, -1), \end{aligned} \tag{2}$$

for all $\mathbf{x} \in \mathcal{X}$. Classifiers are frequently designed to be optimal with respect to the zero-one loss

$$\begin{aligned} L_{0/1}(f, y) &= \frac{1 - sign(yf)}{2} \\ &= \begin{cases} 0, & \text{if } y = sign(f); \\ 1, & \text{if } y \neq sign(f), \end{cases} \end{aligned} \tag{3}$$

where we omit the dependence on $\mathbf{x}$ for notational simplicity. The associated conditional risk is

$$\begin{aligned} C_{0/1}(\eta, f) &= \eta\frac{1 - sign(f)}{2} + (1 - \eta)\frac{1 + sign(f)}{2} \\ &= \begin{cases} 1 - \eta, & \text{if } f \geq 0; \\ \eta, & \text{if } f < 0, \end{cases} \end{aligned} \tag{4}$$

with $\eta(\mathbf{x}) = P_{Y|\mathbf{X}}(1|\mathbf{x})$. This risk is minimized by any predictor f such that

$$\begin{cases} f(\mathbf{x}) > 0 & \text{if } \eta(\mathbf{x}) > \gamma \\ f(\mathbf{x}) = 0 & \text{if } \eta(\mathbf{x}) = \gamma \\ f(\mathbf{x}) < 0 & \text{if } \eta(\mathbf{x}) < \gamma \end{cases} \tag{5}$$

and $\gamma = \frac{1}{2}$. Examples of optimal predictors include $f^* = 2\eta - 1$ and $f^* = \log\frac{\eta}{1-\eta}$. The associated optimal classifier $h^* = sign[f^*]$ is the well known Bayes decision rule, and the associated minimum conditional (zero-one) risk is

$$\begin{aligned} C^*_{0/1}(\eta) = \eta\left(\frac{1}{2} - \frac{1}{2}sign(2\eta - 1)\right) + \\ (1-\eta)\left(\frac{1}{2} + \frac{1}{2}sign(2\eta - 1)\right). \end{aligned} \tag{6}$$

A number of other losses have been proposed in the literature. Popular examples include the exponential loss of boosting, binomial loss of logistic regression, or hinge loss of SVMs. These losses are of the form $L_\phi(f, y) = \phi(yf)$, for different functions $\phi(\cdot)$. The associated conditional risk

$$C_\phi(\eta, f) = \eta\phi(f) + (1-\eta)\phi(-f). \tag{7}$$

is minimized by the predictor

$$f^*_\phi(\eta) = \arg\min_f C_\phi(\eta, f) \tag{8}$$

leading to the minimum conditional risk function $C^*_\phi(\eta) = C_\phi(\eta, f^*_\phi)$.

Conditional risk minimization is closely related to classical probability elicitation in statistics (Savage, 1971). Here, the goal is to find the probability estimator $\hat\eta$ that maximizes the expected reward

$$I(\eta, \hat\eta) = \eta I_1(\hat\eta) + (1-\eta)I_{-1}(\hat\eta), \tag{9}$$

where $I_1(\hat\eta)$ is the reward for prediction $\hat\eta$ when event $y = 1$ holds and $I_{-1}(\hat\eta)$ the corresponding reward when $y = -1$. The functions $I_1(\cdot), I_{-1}(\cdot)$ should be such that the expected reward is maximal when $\hat\eta = \eta$, i.e.

$$I(\eta, \hat\eta) \le I(\eta, \eta) = J(\eta), \ \ \forall\eta \tag{10}$$

with equality if and only if $\hat\eta = \eta$. The following theorem establishes the conditions under which this holds.

Theorem 1. *(Savage, 1971) Let $I(\eta, \hat\eta)$ and $J(\eta)$ be as defined in (9) and (10). Then 1) $J(\eta)$ is convex and 2) (10) holds if and only if*

$$\begin{aligned} I_1(\eta) &= J(\eta) + (1-\eta)J'(\eta) & (11) \\ I_{-1}(\eta) &= J(\eta) - \eta J'(\eta). & (12) \end{aligned}$$

It follows from the theorem that, starting from any convex $J(\eta)$, it is possible to derive $I_1(\cdot), I_{-1}(\cdot)$ so that (10) holds. The next theorem connects this result to the problem of classifier design.

Theorem 2. *(Masnadi-Shirazi & Vasconcelos, 2009) Let $J(\eta)$ be as defined in (10) and f a continuous function. If the following properties hold*

1. *$J(\eta) = J(1-\eta)$,*
2. *f is invertible with symmetry*

$$f^{-1}(-v) = 1 - f^{-1}(v), \tag{13}$$

then the functions $I_1(\cdot)$ and $I_{-1}(\cdot)$ derived with (11) and (12) satisfy the following equalities

$$\begin{aligned} I_1(\eta) &= -\phi(f(\eta)) & (14) \\ I_{-1}(\eta) &= -\phi(-f(\eta)), & (15) \end{aligned}$$

with

$$\phi(v) = -J[f^{-1}(v)] - (1 - f^{-1}(v))J'[f^{-1}(v)]. \tag{16}$$

This theorem connects (9) and (7), establishing a new path for the design of learning algorithms. Rather than specifying a loss ϕ and minimizing $C_\phi(\eta, f)$, so as to obtain whatever optimal predictor f^*_ϕ and minimum expected risk $C^*_\phi(\eta)$ results, it is possible to specify f^*_ϕ and $C^*_\phi(\eta)$ and derive, from (16) with $J(\eta) = -C^*_\phi(\eta)$, the underlying loss ϕ. The only conditions are that $C^*_\phi(\eta) = C^*_\phi(1-\eta)$ and that (13) holds for f^*_ϕ. Note that 1) the symmetry of (13) guarantees that f meets the necessary conditions of (5) for predictor optimality[1], and 2) the condition of $C^*_\phi(\eta) = C^*_\phi(1-\eta)$ encodes the fact that there is no preference for different types of errors[2].

3. Cost sensitive losses and classifier design

In this section we extend the connections between risk minimization and probability elicitation to the cost-sensitive setting. We start by reviewing cost-sensitive losses.

3.1. Cost-sensitive losses

The cost-sensitive extension of the zero-one loss is

$$\begin{aligned} & L_{C_1,C_{-1}}(f, y) = \\ & \frac{1 - sign(yf)}{2}\left(C_1\frac{1 - sign(f)}{2} + C_{-1}\frac{1 + sign(f)}{2}\right) \\ & = \begin{cases} 0, & \text{if } y = sign(f); \\ C_1, & \text{if } y = 1 \text{ and } sign(f) = -1 \\ C_{-1}, & \text{if } y = -1 \text{ and } sign(f) = 1, \end{cases} \end{aligned} \tag{17}$$

[1] see Theorem 4.

[2] the risk, or expected loss, is the same for any two $\mathbf{x}_1$ and $\mathbf{x}_2$ at the same distance from the boundary, where distance is measured is units of posterior probability ($|\eta(\mathbf{x}) - 1/2|$).

where C_1 is the cost of a false negative, or miss, and C_{-1} that of a false positive. The associated conditional risk is

$$\begin{aligned} & C_{C_1,C_{-1}}(\eta, f) = \\ & C_1\eta\frac{1-sign(f)}{2} + (1-\eta)C_{-1}\frac{1+sign(f)}{2} = \\ & = \begin{cases} C_{-1}(1-\eta), & \text{if } f \geq 0; \\ C_1\eta, & \text{if } f < 0, \end{cases} \end{aligned} \tag{18}$$

and is minimized by any predictor that satisfies (5) with $\gamma = \frac{C_{-1}}{C_1+C_{-1}}$. Examples of optimal predictors include $f^*(\eta) = (C_1 + C_{-1})\eta - C_{-1}$ and $f^*(\eta) = \log\frac{\eta C_1}{(1-\eta)C_{-1}}$. The associated optimal classifier $h^* = sign[f^*]$ is the cost-sensitive Bayes decision rule, and the associated minimum conditional (cost-sensitive) risk is

$$\begin{aligned} C^*_{C_1,C_{-1}}(\eta) = C_1\eta\left(\frac{1}{2} - \frac{1}{2}sign\left[f^*(\eta)\right]\right) + \\ C_{-1}(1-\eta)\left(\frac{1}{2} + \frac{1}{2}sign\left[f^*(\eta)\right]\right) \end{aligned} \tag{19}$$

with $f^*(\eta) = (C_1 + C_{-1})\eta - C_{-1}$. To extend the other losses used in machine learning to the cost-sensitive setting, we consider the following set of loss functions

$$\begin{aligned} L_{\phi,C_1,C_{-1}}(f, y) &= \phi_{C_1,C_{-1}}(yf) \\ &= \begin{cases} \phi_1(f), & \text{if } y = 1 \\ \phi_{-1}(-f), & \text{if } y = -1. \end{cases} \end{aligned} \tag{20}$$

The associated conditional risk

$$C_{\phi,C_1,C_{-1}}(\eta, f) = \eta\phi_1(f) + (1-\eta)\phi_{-1}(-f) \tag{21}$$

is minimized by the predictor

$$f^*_{\phi,C_1,C_{-1}}(\eta) = \arg\min_f C_{\phi,C_1,C_{-1}}(\eta, f) \tag{22}$$

leading to the minimum conditional risk

$$\begin{aligned} C^*_{\phi,C_1,C_{-1}}(\eta) &= \eta\phi_1(f^*_{\phi,C_1,C_{-1}}(\eta)) \\ &+ (1-\eta)\phi_{-1}(-f^*_{\phi,C_1,C_{-1}}(\eta)). \end{aligned} \tag{23}$$

3.2. Cost-sensitive learning algorithms

It is currently not known which loss functions $\phi_i(\cdot)$ in (20) best extend the ones used in the design of cost-insensitive algorithms, so as to produce cost-sensitive extensions of boosting, or SVM classifiers. We address this problem by extending the approach of (Masnadi-Shirazi & Vasconcelos, 2009).

Theorem 3. *Let $g(\eta)$ be any invertible function, $J(\eta)$ any convex function, and $\phi_i(\cdot)$ determined by the following steps:*

1. *use (11) and (12) to obtain the $I_1(\eta)$ and $I_{-1}(\eta)$, and let $C_{\phi,C_1,C_{-1}}(\eta, f)$ be defined by (21).*
2. *set $\phi_1(g(\eta)) = -I_1(\eta)$ and $\phi_{-1}(-g(\eta)) = -I_{-1}(\eta)$.*

Then $g(\eta) = f^*_{\phi,C_1,C_{-1}}(\eta)$ if and only if $J(\eta) = -C^*_{\phi,C_1,C_{-1}}(\eta)$.

The theorem shows that any loss with components $\phi_i(\cdot)$ designed according to steps 1. and 2. satisfies (21)-(23), when $g(\eta) = f^*_{\phi,C_1,C_{-1}}(\eta)$ and $J(\eta) = -C^*_{\phi,C_1,C_{-1}}(\eta)$. This implies that it is possible to specify any pair $f^*_{\phi,C_1,C_{-1}}(\eta)$, $C^*_{\phi,C_1,C_{-1}}(\eta)$ and derive the underlying loss. The next question is how to choose the best pair of $f^*_{\phi,C_1,C_{-1}}(\eta)$, and $C^*_{\phi,C_1,C_{-1}}(\eta)$.

The following theorem provides a sufficient condition for the Bayes-optimality of $f^*_{\phi,C_1,C_{-1}}(\eta)$.

Theorem 4. *Any invertible predictor $f(\eta)$ with symmetry*

$$f^{-1}(-v) = \frac{2C_{-1}}{C_1 + C_{-1}} - f^{-1}(v) \tag{24}$$

satisfies the necessary and sufficient conditions for cost-sensitive optimality of (5) with $\gamma = \frac{C_{-1}}{C_1+C_{-1}}$.

Hence, the specification of $f^*_{\phi,C_1,C_{-1}}(\eta)$ as any predictor that satisfies (24) guarantees that the conditional risk is minimized by the cost-sensitive Bayes decision rule. The specification of $C^*_{\phi,C_1,C_{-1}}(\eta)$ determines the risk of the optimal classifier. The goal is to approximate as best as possible the cost-sensitive Bayes risk, given in (19). The next theorem highlights some fundamental properties of this risk.

Theorem 5. *The risk of (19) has the following properties:*

1. *a maximum at $\eta^* = \frac{C_{-1}}{C_1+C_{-1}}$*
2. *symmetry defined by, $\forall\epsilon \in \left[0, \frac{1}{C_1+C_{-1}}\right]$,*

$$C^*\left(\eta^* - C_{-1}\epsilon\right) = C^*\left(\eta^* + C_1\epsilon\right), \tag{25}$$

As noted by the following lemma, property 2. is in fact a generalization of property 1.

Lemma 6. *Any concave function with the symmetry of (25) also has property 1. of Theorem 5.*

Property 1. assigns the largest risk to the locations on the classification boundary. This can be seen as a minimal requirement for consistency of any $C^*_{\phi,C_1,C_{-1}}(\eta)$ with Bayesian decision theory. Enforcing Property 2. further guarantees that the optimal risk has the symmetry of the cost-sensitive Bayes risk. Theorem 5 hence suggests the following risk taxonomy.

Definition 1. *A minimum risk $C^*_{\phi,C_1,C_{-1}}(\eta)$ is of*

1. *Type-I if it satisfies property 1. but not 2. of Theorem 5.*

2. *Type-II if it satisfies both properties 1. and 2.*

Risks of type-II are closer approximations to the cost-sensitive Bayes risk than those of type I.

The combination of Theorems 3-5 leads to a generic procedure for the design of cost-sensitive classification algorithms, consisting of the following steps

1. select a predictor $f^*_{\phi,C_1,C_{-1}}(\eta)$ that satisfies (24).

2. select a concave minimum conditional risk $C^*_{\phi,C_1,C_{-1}}(\eta)$ of type-I or type-II, which reduces to $C^*_\phi(\eta)$ when $C_1 = C_{-1} = 1$.

3. use (11) and (12) with $J(\eta) = -C^*_{\phi,C_1,C_{-1}}(\eta)$ to obtain $I_1(\eta)$ and $I_{-1}(\eta)$.

4. find $\phi_i(\cdot)$ so that $I_1(\eta) = -\phi_1(f^*_{\phi,C_1,C_{-1}}(\eta))$ and $I_{-1}(\eta) = -\phi_{-1}(-f^*_{\phi,C_1,C_{-1}}(\eta))$.

5. derive an algorithm to minimize the conditional risk of (21).

We next illustrate the practical application of this framework by showing that the cost-sensitive exponential loss of (Masnadi-Shirazi & Vasconcelos, 2007) can be derived from a minimal conditional risk of Type-I.

3.3. Cost-sensitive exponential loss

We start by recalling that AdaBoost is based on the loss $\phi(yf) = \exp(-yf)$, for which it can be shown that

$$C^*_\phi(\eta) = \eta\sqrt{\frac{1-\eta}{\eta}} + (1-\eta)\sqrt{\frac{\eta}{1-\eta}}$$
$$\text{and} \quad f^*_\phi = \frac{1}{2}\log\frac{\eta}{1-\eta}. \tag{26}$$

A natural cost-sensitive extension is $f^*_{\phi,C_1,C_{-1}}(\eta) = \frac{1}{C_1+C_{-1}}\log\frac{\eta C_1}{(1-\eta)C_{-1}}$, which is easily shown to satisfy (24). Noting that $C^*_\phi(\eta) = \eta\exp(-f^*_\phi) + (1-\eta)\exp(f^*_\phi)$, suggests the cost-sensitive extension

$$C^*_{\phi,C_1,C_{-1}}(\eta) = \eta\left(\frac{\eta C_1}{(1-\eta)C_{-1}}\right)^{\frac{-C_1}{C_1+C_{-1}}} + (1-\eta)\left(\frac{\eta C_1}{(1-\eta)C_{-1}}\right)^{\frac{C_{-1}}{C_1+C_{-1}}}. \tag{27}$$

This does not have the symmetry of (25) but satisfies property 1. of Theorem 5. Hence, it is a Type-I risk. It is also equivalent to (26) when $C_1 = C_{-1} = 1$. Finally, steps 1. and 2. of Theorem 3 produce the loss

$$\phi_{C_1,C_{-1}}(yf) = \begin{cases} \exp(-C_1 f), & \text{if } y = 1 \\ \exp(C_{-1} f), & \text{if } y = -1 \end{cases} \tag{28}$$

proposed in (Masnadi-Shirazi & Vasconcelos, 2007). The resulting cost-sensitive boosting algorithm currently holds the best performance in the literature.

4. Cost sensitive SVM

We next consider the case of the cost-sensitive SVM. We start by extending the hinge loss, using the framework of the previous section, and then derive the cost-sensitive SVM optimization problem.

4.1. Cost-sensitive hinge-loss

We start by recalling that the SVM minimizes the risk of the hinge loss $\phi(yf) = \lfloor 1 - yf\rfloor_+$, where $\lfloor x\rfloor_+ = \max(x,0)$. This risk is minimized by (Zhang, 2004)

$$f^*_\phi(\eta) = sign(2\eta - 1) \tag{29}$$

leading to the minimum conditional risk

$$C^*_\phi(\eta) = 1 - |2\eta - 1|$$
$$= \eta\lfloor 1 - sign(2\eta-1)\rfloor_+ + (1-\eta)\lfloor 1 + sign(2\eta-1)\rfloor_+.$$

Again, we replace the optimal cost-insensitive predictor by its cost-sensitive counterpart

$$f^*_{\phi,C_1,C_{-1}}(\eta) = sign((C_1 + C_{-1})\eta - C_{-1}). \tag{30}$$

which is easily shown to satisfy (5). This suggests the cost-sensitive minimum conditional risk

$$\begin{aligned} &C^*_{\phi,C_1,C_{-1}}(\eta) = \\ &\eta\lfloor e - d\cdot sign((C_1+C_{-1})\eta - C_{-1})\rfloor_+ + \\ &(1-\eta)\lfloor b + a\cdot sign((C_1+C_{-1})\eta - C_{-1})\rfloor_+, \end{aligned} \tag{31}$$

which can be shown to satisfy (25) if and only if

$$d \geq e \qquad a \geq b \qquad \text{and} \qquad \frac{C_{-1}}{C_1} = \frac{a+b}{d+e}. \tag{32}$$

After steps 1. and 2. of Theorem 3,

$$\phi_{C_1,C_{-1}}(yf) = \begin{cases} \lfloor e - df\rfloor_+, & \text{if } y = 1 \\ \lfloor b + af\rfloor_+, & \text{if } y = -1. \end{cases} \tag{33}$$

This loss has four degrees of freedom, which control the margin and slope of the hinge components associated with the two classes: positive examples are classified with margin $\frac{e}{d}$ and hinge loss slope d, while for negative examples the margin is $\frac{b}{a}$ and slope a.

4.2. Cost-sensitive SVM learning

We consider the case where errors in the positive class are weighted more heavily, leading to the inequalities $\frac{b}{a} \leq \frac{e}{d}$ and $d \geq a$. Choosing $e = d = C_1$ normalizes the margin of positive examples to unity ($\frac{e}{d} = 1$). Selecting $b = 1$ then fixes the scale of the negative component of the hinge loss, leading to $a = 2C_{-1} - 1$. The resulting cost sensitive SVM minimal conditional risk is

$$
\begin{aligned}
&C^*_{\phi,C_1,C_{-1}}(\eta) = \\
&\eta\lfloor C_1 - C_1 \cdot sign((C_1 + C_{-1})\eta - C_{-1})\rfloor_+ + \\
&(1-\eta)\lfloor 1 + (2C_{-1} - 1) \cdot sign((C_1 + C_{-1})\eta - C_{-1})\rfloor_+
\end{aligned} \tag{34}
$$

with $C_{-1} \geq 1$ and $C_1 \geq 2C_{-1} - 1$, so as to satisfy (32). Figure 1 presents plots of (34) and (33), for both $C_1 = 4$, $C_{-1} = 2$ and the cost insensitive case of $C_1 = 1, C_{-1} = 1$ (standard SVM). Note that, for the cost-sensitive SVM, the positive class has a unit margin, while the negative class has a smaller margin of $\frac{1}{3}$. Also, the slope of the positive component of the loss is 4 while the negative component has a smaller slope of 3. In this way, the loss assigns a higher cost to errors in the positive class when the data is not separable, while enforcing a larger margin for positive examples when the data is separable.

Replacing the standard hinge loss with (33) in the standard SVM risk (Moguerza & Munoz, 2006)

$$
\begin{aligned}
&\arg\min_{w,b} \sum_{\{i|y_i=1\}} \lfloor C_1 - C_1(w^T x_i + b)\rfloor_+ \\
&+ \sum_{\{i|y_i=-1\}} \lfloor 1 + (2C_{-1} - 1)(w^T x_i + b)\rfloor_+ + \mu||w||^2,
\end{aligned} \tag{35}
$$

leads to the primal problem

$$
\begin{aligned}
&\arg\min_{w,b} \frac{1}{2}||w||^2 + C\left[\beta \sum_{\{i|y_i=1\}} \xi_i \right. \\
&\qquad\qquad \left. +\lambda \sum_{\{i|y_i=-1\}} \xi_i \right] \\
&\text{s.t. } (w^T x_i + b) \geq 1 - \xi_i; \quad y_i = 1 \\
&(w^T x_i + b) \leq -\kappa + \xi_i; \quad y_i = -1
\end{aligned} \tag{36}
$$

with

$$
\beta = C_1 \qquad \lambda = 2C_{-1} - 1 \qquad \kappa = \frac{1}{2C_{-1} - 1}. \tag{37}
$$

This is a quadratic programming problem similar to that of the standard cost-insensitive SVM with soft margin weight parameter C. In this case, cost-sensitivity is controlled by the parameters β, λ, and κ. The parameter κ is responsible for cost-sensitivity in the separable case. Under the constraints $C_1 \geq 1, C_1 \geq 2C_{-1} - 1$ of a type-II risk, it imposes a smaller margin on negative examples. On the other hand, β and λ control the relative weights of margin violations, assigning more weight to positive violations. This allows control of cost-sensitivity when the data is not separable.

Obviously, this primal problem could be defined through heuristic arguments. However, it would be difficult to justify precise choices for the parameters of (37). Furthermore, the derivation above guarantees that the optimal classifier implements the Bayes decision rule of (5) with $\gamma = \frac{C_{-1}}{C_1 + C_{-1}}$, and its risk is a type-II approximation to the cost-sensitive Bayes risk. No such guarantees would be possible for an heuristic solution.

To obtain some intuition about the cost-sensitive extension, we consider the synthetic problem of Figure 1, where the two classes are linearly separable. The figure shows three separating lines. The green line is an arbitrary separating line that does not maximize the margin. The red line is the standard SVM solution, which has maximum margin and is equally distant from the nearest examples of the two classes. The blue line is the solution of (36) for $C_1 = 4$ and $C_{-1} = 2$ (the C parameter is irrelevant when the data is separable). It is also a maximum margin solution, but trades-off the distance to positive and negative examples so as to enforce a larger positive margin, as specified. Overall, an increase in C_{-1} guarantees a larger positive margin. For a given C_{-1}, increasing C_1 (so that $C_1 \geq 2C_{-1} - 1$) increases the cost of errors on positive examples, enabling control of the miss rate when the classes are not separable.

Finally, the dual and kernelized formulation of the cost sensitive SVM can be obtained with the standard procedures, leading to

$$
\begin{aligned}
&\arg\max_{\alpha_i} \sum_i \alpha_i \left(\frac{y_i + 1}{2} - \frac{y_i - 1}{2(2C_{-1} - 1)}\right) \\
&\qquad -\frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\
&\text{s.t.} \qquad \sum_i \alpha_i y_i = 0 \\
&\qquad 0 \leq \alpha_i \leq CC_1; \qquad y_i = 1 \\
&\qquad 0 \leq \alpha_i \leq C(2C_{-1} - 1); \quad y_i = -1.
\end{aligned} \tag{38}
$$

This reduces to the standard SVM dual when $C_1 = C_{-1} = 1$. Note that the derivation of the cost-sensitive SVM from a suitable loss function leads to an algorithm that performs regardless of the separability of the data and slack penalty, unlike the previous BM-SVM and BP-SVM algorithms. The improved performance of CS-SVM on real world data sets is demonstrated in the next section.

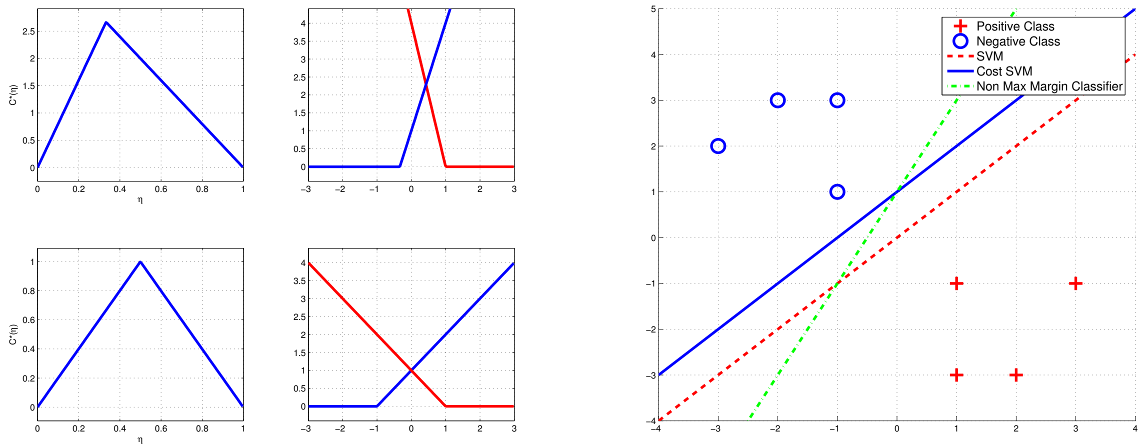

Figure 1. Left: concave $C^*_{\phi,C_1,C_{-1}}(\eta)$ function and corresponding cost sensitive SVM loss function, top: $C_1 = 4$, $C_{-1} = 2$, bottom: $C_1 = C_{-1} = 1$. Right: linearly separable cost sensitive SVM.

5. Experimental results

The performance of the CS-SVM was evaluated with two sets of experiments. The first was based on ten binary UCI data sets (Newman et al., 1998): Pima-diabetes, breast cancer diagnostic, breast cancer prognostic, original Wisconsin breast cancer, liver disorder, sonar, echo-cardiogram, Cleveland heart disease, tic-tac-toe and Haberman's survival. The goal was to learn the SVM of lowest total error rate, given a target detection rate. In all cases, leave one out cross validation was used to find the best cost estimate. We considered detection rates between 80% and 95%, with increments of 2.5%, and set C, C_1, C_{-1} and b (SVM threshold) for each method so as to achieve the smallest false positive rate on the validation set. The total error was computed for each detection rate, and the mean of these errors is reported in Table-2. Results are reported for the proposed CS-SVM, the BM-SVM (Karakoulas & Shawe-Taylor, 1999) and the BP-SVM (Bach et al., 2006; Lin et al., 2002; Davenport et al., 2006; Wu & Srihari, 2003; Chang & Lin, 2001). While the table confirms the previous observation that the BP-SVM outperforms the BM-SVM (Bach et al., 2006; Lin et al., 2002; Davenport et al., 2006; Wu & Srihari, 2003; Chang & Lin, 2001), none of them matches the CS-SVM. This is most interesting given the fact that CS-SVM has the same computational complexity and number of tuning parameters as the BP-SVM. Overall, CS-SVM has the smallest error on 7 of the 10 datasets, sometimes by a very substantial margin. CS-SVM and BP-SVM have equal error on 2 datasets, and BP and BM-SVMs have a slight advantage on Wisconsin.

The second set of experiments was based on the German Credit data set (Geibel et al., 2004; Newman et al., 1998). This dataset has 700 examples of good credit customers and 300 examples of bad credit customers. Each example is described by 24 attributes, and the goal is to identify bad costumers, to be denied credit. This data set is particularly interesting for cost-sensitive learning because it provides a cost matrix for the different types of errors. Classifying a good credit customer as bad (a false-positive) incurs a loss of 1. Classifying a bad credit customer as good (a miss) incurs a loss of 5. Hence, on this dataset, the leave one out cross validation of CS-SVM and BP-SVM parameters was subject to the constraint $\frac{C_1}{C_{-1}} = 5$. A cost insensitive SVM was also trained. Table 1 presents the loss achieved by each method. Note that BP-SVM does not produce any improvement with respect to the cost insensitive SVM. On the other hand, the loss achieved with CS-SVM is 328\$ smaller, i.e. a substantial reduction of cost by 37.36%.

Table 1. Total loss in \$ for each method on the German Credit dataset.

Method	CS-SVM	BP-SVM	SVM
Loss \$	550\$	878\$	878\$

6. Conclusion

In this work, we have extended the recently introduced probability elicitation view of loss function design to the cost sensitive classification problem. This extension was applied to the SVM problem, so as to produce a cost-sensitive hinge loss function. A cost-sensitive SVM learning algorithm was then derived, as the minimizer of the associated risk. Unlike previous SVM algorithms, the one

Table 2. mean error for each UCI data set and cost sensitive SVM method.

Dataset	Survive	Liver	Echo	Pima	Wisc	Tic	Heart	Diag	Prag	Sonar
CS-SVM	**195.8**	**163.8**	**40**	**313.2**	33.2	536	**68.4**	33.8	**107.2**	**65.6**
BP-SVM	199.6	167.2	43	416	32.8	536	69.4	33.8	115.2	75.2
BM-SVM	201.8	169.2	45	416	32.8	538	73.2	33.8	126	76.4

now proposed enforces cost sensitivity for both separable and non-separable training data, enforcing a larger margin for the preferred class, independent of the choice of slack penalty. It also offers guarantees of optimality, namely classifiers that implement the cost-sensitive Bayes decision rule and approximate the cost-sensitive Bayes risk. Empirical evidence confirms its superior performance, when compared to previous methods.

References

Akbani, Rehan, Kwek, Stephen, and Japkowicz, Nathalie. Applying support vector machines to imbalanced datasets. In *European Conference on Machine Learning (ECML)*, pp. 39–50, 2004.

Amari, S. and Wu, S. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12(6):783–789, 1999.

Bach, Francis R., Heckerman, David, and Horvitz, Eric. Considering cost asymmetry in learning classifiers. *The Journal of Machine Learning Research*, 7:1713–1741, 2006.

Chang, Chih-Chung and Lin, Chih-Jen. *LIBSVM: a library for support vector machines*, 2001.

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

Cortes, Corinna and Vapnik, Vladimir. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

Davenport, M.A., Baraniuk, R.G., and Scott, C.D. Controlling false alarms with support vector machines. In *ICASSP*, 2006.

Elkan, C. The foundations of cost-sensitive learning. In *Joint Conference on Artificial Intelligence*, 2001.

Geibel, Peter, Brefeld, Ulf, and Wysotzki, Fritz. Perceptron and svm learning with generalized cost models. *Intelligent Data Analysis*, 8:439–455, 2004.

Karakoulas, G. and Shawe-Taylor, J. optimizing classifiers for imbalanced training sets. In *NIPS*, 1999.

Kubat, Miroslav and Matwin, Stan. Addressing the curse of imbalanced training sets: One-sided selection. In *ICML*, pp. 179–186, 1997.

Lin, Yi, Lee, Yoonkyung, and Wahba, Grace. Support vector machines for classification in nonstandard situations. *Machine Learning*, 46:191–202, 2002.

Masnadi-Shirazi, Hamed and Vasconcelos, Nuno. Asymmetric boosting. In *ICML*, 2007.

Masnadi-Shirazi, Hamed and Vasconcelos, Nuno. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In *NIPS*, pp. 1049–1056, 2009.

Moguerza, Javier M. and Munoz, Alberto. Support vector machines with applications. *Statistical Science*, 21:322, 2006.

Newman, D.J., Hettich, S., Blake, C.L., and Merz, C.J. UCI repository of machine learning databases, 1998.

Platt, J. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In *Adv. in Large Margin Classifiers*, 2000.

Savage, Leonard J. The elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66:783–801, 1971.

Wu, G. and Chang, E. Adaptive feature-space conformal transformation for imbalanced data learning. In *ICML*, pp. 816–823, 2003.

Wu, G. and Chang, E. Kba: kernel boundary alignment considering imbalanced data distribution. *IEEE Transactions on Knowledge and Data Engineering*, 17:786–795, 2005.

Wu, X. and Srihari, R. New ν-support vector machines and their sequential minimal optimization. In *ICML*, 2003.

Zhang, Tong. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 2004.