# Cost-sensitive Support Vector Machines

**Arya Iranmehr** AIRANMEHR@UCSD.EDU
*Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92039*
**Hamed Masnadi-Shirazi** HMASNADI@SHIRAZU.AC.IR
*Department of Electrical Engineering, Shiraz University, Shiraz, Iran*
**Nuno Vasconcelos** NUNO@UCSD.EDU
*Statistical Visual Computing Laboratory, University of California, San Diego, La Jolla, CA 92039*

## Abstract

Many machine learning applications involve imbalance class prior probabilities, multi-class classification with many classes, or "cost-sensitive" classification. In such domains, each class (or in some cases, each sample) requires an special treatment.

In this paper, we use a constructive procedure to extend SVM's standard loss function to optimize classifier with respect to class imbalance or class costs. By drawing connections between risk minimization and probability elicitation, we show that the resulting classifier guarantees Bayes consistency. We further analyze the primal and the dual objective functions and derive the objective function in a regularized risk minimization framework. Finally, we extend the classifier to the with cost-sensitive learning with example dependent costs. We perform experimental analysis on class imbalance, cost-sensitive with given class and example costs and show that proposed algorithm provides superior generalization performance, compared to the conventional methods.

**Keywords:** Cost Sensitive Learning, Classification, Class Imbalance, SVM, Bayes Consistency

## 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 and 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 and Matwin (1997); Chawla et al. (2002); Akbani et al. (2004); Geibel et al. (2004); Zadrozny et al. (2003); Tang et al. (2009); Köknar-Tezel and Latecki (2009); Wang et al. (2012a); Mathew et al. (2017); Zeng and Gao (2009); Wang et al. (2012b). 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. Active learning based methods have also been proposed to train the SVM algorithm on the informative instances, instances which are close to the hyperplane Ertekin et al. (2007). The second class of approaches Amari and Wu (1999); Wu and Chang (2003, 2005) involve 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 and Shawe-Taylor (1999). Under Bayesian decision theory, this would be the optimal 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), 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 and Srihari (2003); Chang and Lin (2011). 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

$$\begin{aligned} \underset{w,b,\xi}{\text{argmin}} \quad & \frac{1}{2}||w||^2 + C\left[C_1 \sum_{\{i|y_i=1\}} \xi_i + C_{-1} \sum_{\{i|y_i=-1\}} \xi_i\right] \\ \text{subject to} \quad & y_i(w^T x + b) \geq 1 - \xi_i. \end{aligned} \tag{1}$$

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 $\xi_i$ 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 and Vasconcelos (2008). 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.

Since CS-SVM is implemented in the dual, cost-sensitive learning in the dual should be studied more closely. We show that cost-sensitive learning in the dual appears as regularization and changing the constraint's upper bounds which stem from sensitivity analysis. These connections are considered under cost-sensitive learning and imbalanced data learning.

Moreover, we show that in the cost-sensitive and imbalanced data settings, the priors and costs should be incorporated in the performance measure. We propose minimum expected (cost-sensitive) risk as a cost sensitive performance metric and demonstrate its connections to the ROC curve. For the case of unknown costs, we introduce a robust measure which reflects the performance of the classifier under a given tolerance of false-positive or false-negative errors.

The paper is organized as follows. Section 2 briefly reviews the probability elicitation view of loss function design Masnadi-Shirazi and Vasconcelos (2008). 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. In section 5, the dual problem of CS-SVM is thoroughly evaluated in the sense of regularization and sensitivity analysis. Section 6 presents an extension of CS-SVM for problems with example-dependent costs. Section 7 proposes minimum cost sensitive risk as a standard measure for examining classifier performance in the cost-sensitive and imbalanced data setting. Finally, Section 8 presents an experimental evaluation that demonstrates improved performance of the proposed cost sensitive SVM over previous methods.

## 2. Bayes consistent classifier design

The goal of classification is to map feature vectors $\mathbf{x} \in \mathcal{X}$ to class labels $y \in \{-1, 1\}$. From a statistical viewpoint, the feature vectors and class labels are drawn from probability distributions $P_{\mathbf{X}}(\mathbf{x})$ and $P_Y(y)$ respectively. In terms of functions, we write a classifier as $h(\mathbf{x}) = sign[p(\mathbf{x})]$, where the function $p : \mathcal{X} \to \mathbb{R}$ is denoted as the classifier predictor. Given a non-negative function $L(p(\mathbf{x}), y)$ that assigns a loss to each $(p(\mathbf{x}), y)$ pair, the classifier is considered optimal if it minimizes the expected loss $R = E_{\mathbf{X},Y}[L(p(\mathbf{x}), y)]$, also known as the risk. Minimizing the risk, is itself equivalent to minimizing the conditional risk

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

for all $\mathbf{x} \in \mathcal{X}$. It is discerning to write the predictor function $p(\mathbf{x})$ as a composition of two functions $p(\mathbf{x}) = f(\eta(\mathbf{x}))$, where $\eta(\mathbf{x}) = P_{Y|\mathbf{X}}(1|\mathbf{x})$ is the posterior probability , and $f : [0, 1] \to \mathbb{R}$ is denoted as the *link function*. This provides a valuable connection to the Bayes decision rule. A loss is considered Bayes consistent when its associated risk is minimized by the BDR. For example the zero-one loss can be written as

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

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

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

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 thus proving that the zero-one loss is Bayes consistent. Finally, 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 handful of other losses have been shown to be Bayes consistent. These include the exponential loss used in boosting classifiers Friedman et al. (2000), logistic loss of logistic regression Friedman et al. (2000); Zhang (2004), or the hinge loss of SVMs Zhang (2004). These losses are of the form $L_\phi(f,y) = \phi(yf)$ for different functions $\phi(\cdot)$ and are known as *margin losses*. Margin losses assign a non-zero penalty to small positive $yf$, encouraging the creation of a margin. The resulting large-margin classifiers have better generalization than those produced by the zero-one loss or other losses that do not enforce a margin Vapnik (1998). For a margin loss, the conditional risk is simply

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

The conditional risk is minimized by the predictor

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

and the minimum conditional risk is $C^*_\phi(\eta) = C_\phi(\eta, f^*_\phi)$.

Recently, a generative formula for the derivation of novel Bayes consistent loss functions has been presented in Masnadi-Shirazi and Vasconcelos (2008) relying on classical probability elicitation in statistics Savage (1971). Comparable to risk minimization, in probability elicitation, 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 predicting $\hat{\eta}$ when event $y = 1$ holds and $I_{-1}(\hat{\eta})$ the corresponding reward when $y = -1$. The functions $I_1(\cdot), I_{-1}(\cdot)$ are such that the expected reward is maximal when $\hat{\eta} = \eta$, i.e.

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

with equality if and only if $\hat{\eta} = \eta$.

**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*

$$I_1(\eta) = J(\eta) + (1-\eta)J'(\eta) \tag{11}$$

$$I_{-1}(\eta) = J(\eta) - \eta J'(\eta). \tag{12}$$

The theorem states that $I_1(\cdot), I_{-1}(\cdot)$ can be derived such that (10) holds by applying an appropriate convex $J(\eta)$. This primary theorem was used in Masnadi-Shirazi and Vasconcelos (2008) to establish the following for margin loss functions.

**Theorem 2** *Masnadi-Shirazi and Vasconcelos (2008) 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 provides a generative path for designing Bayes consistent margin loss functions for classification. Specifically, any convex symmetric function $J(\eta) = -C^*_\phi(\eta)$ and invertible function $f^{-1}$ satisfying (13) can be used in equation (16) to derive a novel Bayes consistent loss function $\phi(v)$. This is in contrast to previous approaches which require guessing a loss function $\phi(v)$ and checking that it is Bayes consistent by minimizing $C_\phi(\eta, f)$, so as to obtain whatever optimal predictor $f^*_\phi$ and minimum expected risk $C^*_\phi(\eta)$ results Zhang (2004) or methods that restrict the loss function to being convex, differentiable at zero, and have negative derivative at the origin Bartlett et al. (2006).

## 3. Cost sensitive Bayes consistent classifier design

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

### 3.1 Cost-sensitive zero-one loss

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

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

where $C_1$ is the cost of a false negative 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 $h^* = sign[f^*]$ implements the cost-sensitive Bayes decision rule, and the associated minimum conditional (cost-sensitive) risk is

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

with $f^*(\eta) = (C_1 + C_{-1})\eta - C_{-1}$. We show that the minimum cost sensitive zero-one risk is equivalent to the *minimum cost sensitive Bayes error*.

**Theorem 3** *The minimum risk associated with the cost sensitive zero-one loss is equal to the minimum cost sensitive Bayes error.*

**Proof**

$$R^*_{C_1,C_{-1}} = E_X[C^*_{C_1,C_{-1}}(\eta)] = \int P(x)C^*_{C_1,C_{-1}}(P(1|x))dx = \tag{20}$$

$$\int_{P(1|x)\geq\gamma}(\frac{P(x|1)+P(x|-1)}{2})(C_{-1}(1-\frac{P(x|1)}{P(x|1)+P(x|-1)}))dx + \tag{21}$$

$$\int_{P(1|x)<\gamma}(\frac{P(x|1)+P(x|-1)}{2})(C_1(\frac{P(x|1)}{P(x|1)+P(x|-1)}))dx = \tag{22}$$

$$\frac{1}{2}\int_{P(1|x)\geq\gamma}C_{-1}P(x|-1)dx + \frac{1}{2}\int_{P(1|x)<\gamma}C_1P(x|1)dx = \tag{23}$$

$$\frac{1}{2}(C_{-1}\epsilon_1^{\gamma} + C_1\epsilon_2^{\gamma}) = \epsilon_{C_1,C_{-1}} \tag{24}$$

where $\epsilon_2^{\gamma}$ and $\epsilon_1^{\gamma}$ are the miss rate and false positive rate associated with the cost sensitive threshold $\gamma$ and $\epsilon_{C_1,C_{-1}}$ is the cost sensitive Bayes error rate. We have also assumed, without loss of generality, that the prior probabilities are equal. ■

The next theorem highlights some fundamental properties of the minimum conditional cost-sensitive zero-one risk.

**Theorem 4** *The risk of (19) has the following properties:*

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

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

**Proof** Note that (19) can be written as

$$C^*_{C_1,C_{-1}}(\eta) = \begin{cases} C_{-1}(1-\eta), & \text{if } f^* \geq 0; \\ C_1\eta, & \text{if } f^* < 0, \end{cases} \tag{26}$$

The two lines $C_{-1}(1-\eta)$ and $C_1\eta$ intersect and form the maximum at $\eta = \frac{C_{-1}}{C_1+C_{-1}}$.

When $\epsilon = 0$ we have the trivial case of $C^*\left(\frac{C_{-1}}{C_1+C_{-1}}\right) = C^*\left(\frac{C_{-1}}{C_1+C_{-1}}\right)$.

When $0 < \epsilon \leq \frac{1}{C_1+C_{-1}}$ we have $\eta = \frac{C_{-1}}{C_1+C_{-1}} - C_{-1}\epsilon < \frac{C_{-1}}{C_1+C_{-1}}$ in which case from (5), $f^* < 0$ and

$$C^*_{C_1,C_{-1}}(\eta) = C_1\eta = C_1\left(\frac{C_{-1}}{C_1+C_{-1}} - C_{-1}\epsilon\right) = \frac{C_1C_{-1}}{C_1+C_{-1}} - C_1C_{-1}\epsilon \tag{27}$$

When $0 < \epsilon \leq \frac{1}{C_1+C_{-1}}$ we also have $\eta = \frac{C_{-1}}{C_1+C_{-1}} + C_1\epsilon > \frac{C_{-1}}{C_1+C_{-1}}$ in which case from (5), $f^* > 0$ and

$$C^*_{C_1,C_{-1}}(\eta) = C_{-1}(1-\eta) = C_{-1}\left(1 - \frac{C_{-1}}{C_1+C_{-1}} - C_1\epsilon\right) = \frac{C_1C_{-1}}{C_1+C_{-1}} - C_1C_{-1}\epsilon \tag{28}$$

Thus proving that

$$C^*_{C_1,C_{-1}}\left(\frac{C_{-1}}{C_1+C_{-1}} - C_{-1}\epsilon\right) = C^*_{C_1,C_{-1}}\left(\frac{C_{-1}}{C_1+C_{-1}} + C_1\epsilon\right) = \frac{C_1C_{-1}}{C_1+C_{-1}} - C_1C_{-1}\epsilon \tag{29}$$

■

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

**Lemma 1** *Any concave function with the symmetry of (25) also has property 1. of Theorem 4.*

**Proof** Taking the derivative of (25) at $\epsilon = 0$ leads to

$$C^{*'}\left(\frac{C_{-1}}{C_1+C_{-1}}\right)(-C_{-1}) = C^{*'}\left(\frac{C_{-1}}{C_1+C_{-1}}\right)(C_1) \tag{30}$$

which is satisfied only when $C^{*'}\left(\frac{C_{-1}}{C_1+C_{-1}}\right) = 0$. Given that $C^*$ is a concave function, $C^*$ is maximum at $\eta = \frac{C_{-1}}{C_2+C_{-1}}$. ■

### 3.2 Cost-sensitive Bayes consistent margin losses

We extend the other losses used in machine learning to the cost-sensitive paradigm by introducing the following set of margin loss function

$$\begin{aligned} L_{\phi,C_1,C_{-1}}(f,y) &= \phi_{C_1,C_{-1}}(yf) \\ &= \begin{cases} \phi_1(f), & \text{if } y = 1 \\ \phi_{-1}(-f), & \text{if } y = -1. \end{cases} \end{aligned} \tag{31}$$

The associated conditional risk is

$$C_{\phi,C_1,C_{-1}}(\eta,f) = \eta\phi_1(f) + (1-\eta)\phi_{-1}(f) \tag{32}$$

and is minimized by the predictor

$$f^*_{\phi,C_1,C_{-1}}(\eta) = \arg\min_f C_{\phi,C_1,C_{-1}}(\eta,f). \tag{33}$$

This leads to the minimum conditional risk

$$
\begin{aligned}
C^*_{\phi,C_1,C_{-1}}(\eta) \quad &= \eta\phi_1(f^*_{\phi,C_1,C_{-1}}(\eta)) \\
&+ (1-\eta)\phi_{-1}(-f^*_{\phi,C_1,C_{-1}}(\eta)).
\end{aligned} \tag{34}
$$

Similar to the cost insensitive case, our choice of $\phi_i(\cdot)$ in (31) cannot be arbitrary and we require certain properties for the loss function. These desirable properties are addressed by extending the approach of Masnadi-Shirazi and Vasconcelos (2008).

**Theorem 5** *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 (32).*

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 $J(\eta) = -C^*_{\phi,C_1,C_{-1}}(\eta)$.*

**Proof** From 1. and Theorem 1, it follows that

$$
\eta I_1(\hat{\eta}) + (1-\eta)I_1(\hat{\eta})
$$

has maximum value $J(\eta)$, when $\hat{\eta} = \eta$. From 2. the same holds for

$$
-\eta\phi_1(g(\hat{\eta})) - (1-\eta)\phi_{-1}(-g(\hat{\eta}))
$$

and

$$
J(\eta) = -\eta\phi_1(g(\eta)) - (1-\eta)\phi_{-1}(-g(\eta)).
$$

It follows from (32)-(34) that, $g(\eta) = f^*_{\phi,C_1,C_{-1}}(\eta)$ if and only if $J(\eta) = -C^*_{\phi,C_1,C_{-1}}(\eta)$. ■

The theorem provides a generative method for designing the loss functions $\phi_i(\cdot)$ starting from any pair of invertible function $g(\eta)$ and convex function $J(\eta)$. The resulting loss function will satisfy (32)-(34), when $g(\eta) = f^*_{\phi,C_1,C_{-1}}(\eta)$ and $J(\eta) = -C^*_{\phi,C_1,C_{-1}}(\eta)$.

What remains to be answered is how to choose $f^*_{\phi,C_1,C_{-1}}(\eta)$, and $C^*_{\phi,C_1,C_{-1}}(\eta)$ so as to ensure cost sensitive Bayes consistency. The following theorem provides a sufficient condition on $f^*_{\phi,C_1,C_{-1}}(\eta)$ for the Bayes optimality of the loss function.

**Theorem 6** *Any invertible predictor $f(\eta)$ with symmetry*

$$
f^{-1}(-v) = \frac{2C_{-1}}{C_1 + C_{-1}} - f^{-1}(v) \tag{35}
$$

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

**Proof** Assume that $f(\eta) = v$ is monotonically increasing. Note that $f^{-1}(0) = \frac{C_{-1}}{C_1+C_{-1}}$ which along with $\eta = f^{-1}(v)$ leads to $f(\frac{C_{-1}}{C_1+C_{-1}}) = 0$. If $\eta > \frac{C_{-1}}{C_1+C_{-1}}$ then from (35) we have $f^{-1}(-v) < \frac{C_{-1}}{C_1+C_{-1}}$, applying (35) again it follows that $f(\eta) > \frac{C_{-1}}{C_1+C_{-1}}$. Similarly, if $\eta < \frac{C_{-1}}{C_1+C_{-1}}$ then $f(\eta) < \frac{C_{-1}}{C_1+C_{-1}}$. ∎

In other words, any predictor $f^*_{\phi,C_1,C_{-1}}(\eta)$ that satisfies (35) will be guaranteed to have a conditional risk that is minimized by the cost-sensitive Bayes decision rule.

What remains to be discussed is how to specify $C^*_{\phi,C_1,C_{-1}}(\eta)$ which will determine the risk of the optimal classifier. The goal is to approximate the minimum conditional cost-sensitive zero-one risk (minimum cost sensitive Bayes risk) given in (19) as best as possible so as to achieve the minimum cost sensitive Bayes error. This is formally presented in the following theorem

**Theorem 7** *The minimum risk of any cost sensitive loss in the form of (31) and derived from Theorem 5 can be made to be arbitrarily close, in the expectation, to the minimum cost sensitive Bayes error by choosing the minimum conditional risk of the loss to be arbitrarily close to the minimum conditional risk of the cost sensitive zero-one loss function.*

**Proof**

$$R^*_{\phi,C_1,C_{-1}} - \epsilon_{C_1,C_{-1}} = R^*_{\phi,C_1,C_{-1}} - R^*_{C_1,C_{-1}} = \tag{36}$$
$$E_X[C^*_{\phi,C_1,C_{-1}}] - E_X[C^*_{C_1,C_{-1}}] = E_X[C^*_{\phi,C_1,C_{-1}} - C^*_{C_1,C_{-1}}] \tag{37}$$

Where we have used Theorem 3 for the first equality. ∎

While Theorem 7 says that the true measure for determining $C^*_{\phi,C_1,C_{-1}}$ is the expectation of (37), Theorem 4 suggests a simpler rule of thumb for selecting $C^*_{\phi,C_1,C_{-1}}$. Property 1. assigns the largest risk to the locations on the classification boundary and requiring this property for $C^*_{\phi,C_1,C_{-1}}$ would be vital. Also, enforcing Property 2. further guarantees that the optimal risk has the symmetry of the minimum cost-sensitive Bayes risk.

**Definition 2** *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 4.*

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

Risks of type-II are generally closer approximations to the cost-sensitive Bayes risk than those of type I. Although, strictly speaking the true measure is the expectation of (37).

The combination of Theorems 4-7 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 (35).

2. select a concave minimum conditional risk using the measure of (37) or, as a simpler rule of thumb alternative, 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 (32).

We next illustrate the practical application of this framework by showing that the cost-sensitive exponential loss of Masnadi-Shirazi and 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

$$\begin{aligned} C^*_\phi(\eta) &= \eta\sqrt{\frac{1-\eta}{\eta}} + (1-\eta)\sqrt{\frac{\eta}{1-\eta}} \\ \text{and} \quad f^*_\phi &= \frac{1}{2}\log\frac{\eta}{1-\eta}. \end{aligned} \tag{38}$$

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 (35). Noting that $C^*_\phi(\eta) = \eta\exp(-f^*_\phi) + (1-\eta)\exp(f^*_\phi)$, suggests the cost-sensitive extension

$$\begin{aligned} C^*_{\phi,C_1,C_{-1}}(\eta) &= \eta\left(\frac{\eta C_1}{(1-\eta)C_{-1}}\right)^{\frac{-C_1}{C_1+C_{-1}}} + \\ &\quad (1-\eta)\left(\frac{\eta C_1}{(1-\eta)C_{-1})}\right)^{\frac{C_{-1}}{C_1+C_{-1}}}. \end{aligned} \tag{39}$$

This does not have the symmetry of (25) but satisfies property 1. of Theorem 4. Hence, it is a Type-I risk. It is also equivalent to (38) when $C_1 = C_{-1} = 1$. Finally, steps 1. and 2. of Theorem 5 produce the loss

$$\phi_{C_1,C_{-1}}(yf) = \begin{cases} \exp(-C_1 f), & \text{if } y = 1 \\ \exp(C_{-1} f), & \text{if } y = -1 \end{cases} \tag{40}$$

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

## 4. Cost sensitive SVM

Next we extend the hinge loss used in SVMs using the cost sensitive framework established in the previous section. The cost sensitive SVM optimization problem is also derived.

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

$$f^*_\phi(\eta) = sign(2\eta - 1) \tag{41}$$

resulting in the minimum conditional risk

$$\begin{aligned} C^*_\phi(\eta) &= 1 - |2\eta - 1| \\ &= \eta\lfloor 1 - sign(2\eta-1)\rfloor_+ + (1-\eta)\lfloor 1 + sign(2\eta - 1)\rfloor_+. \end{aligned}$$

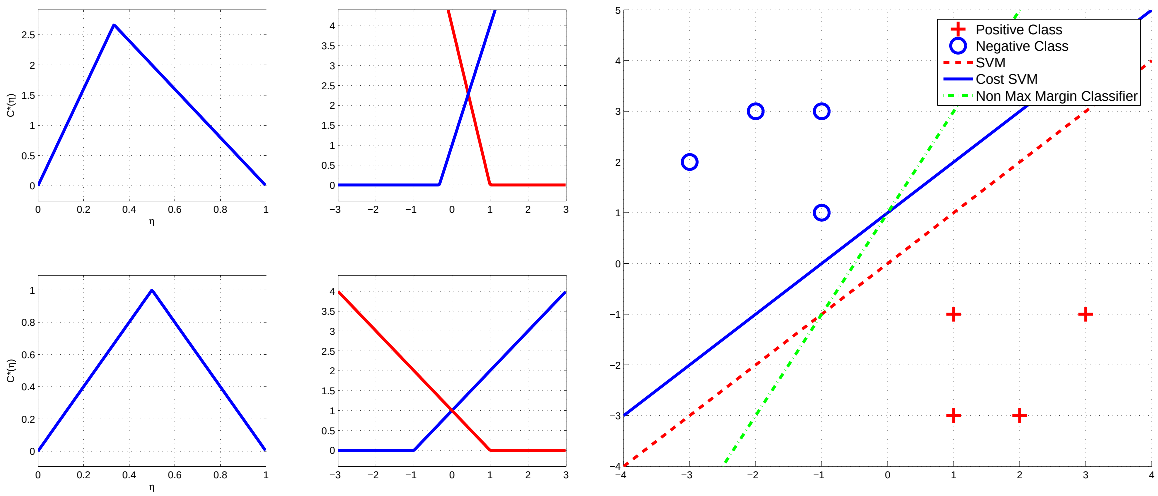



Figure 1: Left: concave $C^*_{\phi,C_1,C_{-1}}(\eta)$ function and corresponding cost sensitive SVM loss function, top: $C_1 = 4$, $C_{-1} = 2$, bottom: $C_1 = C_{-1} = 1$. Right: linearly separable cost sensitive SVM.

We follow the generic procedure and replace the optimal cost-insensitive predictor by its cost-sensitive counterpart

$$f^*_{\phi,C_1,C_{-1}}(\eta) = sign((C_1 + C_{-1})\eta - C_{-1}). \tag{42}$$

which can be directly shown to satisfy (5). This suggests choosing the cost-sensitive minimum conditional risk

$$\begin{aligned} C^*_{\phi,C_1,C_{-1}}(\eta) = \quad & \eta\lfloor e - d \cdot sign((C_1 + C_{-1})\eta - C_{-1})\rfloor_+ + \\ & (1-\eta)\lfloor b + a \cdot sign((C_1 + C_{-1})\eta - C_{-1})\rfloor_+, \end{aligned} \tag{43}$$

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

$$d \geq e \qquad a \geq b \quad \text{and} \quad \frac{C_{-1}}{C_1} = \frac{a+b}{d+e}. \tag{44}$$

The hinge loss minimum conditional risk satisfies the conditions of a Type-II loss function and is also a close approximation of the zero-one minimum conditional risk under the criteria of Theorem -7.

After steps 1. and 2. of Theorem 5,

$$\phi_{C_1,C_{-1}}(yf) = \begin{cases} \lfloor e - df\rfloor_+, & \text{if } y = 1 \\ \lfloor b + af\rfloor_+, & \text{if } y = -1. \end{cases} \tag{45}$$

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 $a$.

### 4.1 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 loss function is

$$\phi_{C_1,C_{-1}}(yf) = 1_{\{y=1\}} C_1 \lfloor 1 - yf \rfloor_+ + 1_{\{y=-1\}} \lfloor 1 - (2C_{-1} - 1) yf \rfloor_+ \tag{46}$$

and the cost sensitive SVM minimal conditional risk is

$$\begin{aligned} &C^*_{\phi,C_1,C_{-1}}(\eta) = \\ &\eta \lfloor C_1 - C_1 \cdot sign((C_1 + C_{-1})\eta - C_{-1}) \rfloor_+ + \\ &(1-\eta) \lfloor 1 + (2C_{-1} - 1) \cdot sign((C_1 + C_{-1})\eta - C_{-1}) \rfloor_+ \end{aligned} \tag{47}$$

with $C_{-1} \geq 1$ and $C_1 \geq 2C_{-1} - 1$, so as to satisfy (44). Figure 1 presents plots of (47) and (46), 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 (45) in the standard SVM risk Moguerza and Munoz (2006)

$$\underset{w,b}{\operatorname{argmin}} \sum_{\{i|y_i=1\}} \lfloor C_1 - C_1(w^T x_i + b) \rfloor_+ + \sum_{\{i|y_i=-1\}} \lfloor 1 + (2C_{-1} - 1)(w^T x_i + b) \rfloor_+ + \frac{1}{2C} ||w||^2, \tag{48}$$

leads to the primal problem

$$\begin{aligned} \underset{w,b,\xi_i}{\operatorname{argmin}} \quad & \frac{1}{2}||w||^2 + C \left[ C_1 \sum_{\{i|y_i=1\}} \xi_i + \frac{1}{\kappa} \sum_{\{i|y_i=-1\}} \xi_i \right] \\ \text{subject to} \quad & (w^T x_i + b) \geq 1 - \xi_i; \qquad y_i = 1 \\ & (w^T x_i + b) \leq -\kappa + \xi_i; \quad y_i = -1 \end{aligned} \tag{49}$$

with

$$\kappa = \frac{1}{2C_{-1} - 1}, \qquad 0 < \kappa \leq 1 \leq \frac{1}{\kappa} \leq C_1. \tag{50}$$

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 $C_1, \frac{1}{\kappa}$, and $\kappa$. The parameter $\kappa$ is responsible for cost-sensitivity in the separable case. Under the constraints $C_{-1} \geq 1$, $C_1 \geq 2C_{-1} - 1$, ($0 < \kappa \leq 1 \leq \frac{1}{\kappa} \leq C_1$), of a type-II risk, it imposes a smaller margin on negative examples. On the other hand, $C_1$ and $\frac{1}{\kappa}$ 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 (50). 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 (49) 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 $C_{-1}$ (decrease in $\kappa$) 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.

We note that for the separable case, a limited level of cost sensitive performance can be achieved using the BP-SVM formulation of (1) along with a small weight parameter $C$ ($C < \frac{1}{2}$), but a small $C$ is undesirable in general as it leads to an under trained model with training errors even when the data is separable. The CS-SVM formulation, on the other hand, provides a maximum margin solution regardless of the chosen weight parameter $C$. The CS-SVM is preferable even in the inseparable case because increasing the weight parameter $C$, in an attempt to reduce training error, inevitably leads to over training in the BP-SVM formulation. This is not necessarily the case for the CS-SVM formulation which allows a decrease of the margin of the negative samples (through an appropriate choice of $\kappa$) and a relative increase in the margin of the positive samples, independent of the weight parameter $C$ and does not lead to over training. In other words, unlike the BP-SVM formulation, the CS-SVM does not simply over train on the positive class, it maximizes the margin on this class. This can also be seen, with added clarity, in the dual CS-SVM formulation which is discussed in the next section.

## 5. Cost-sensitive SVM in the dual

The dual and kernelized formulation of the CS-SVM of (49) can be derived as

$$\begin{aligned}
\underset{\alpha}{\operatorname{argmax}} \quad & \sum_i \alpha_i \left( \frac{y_i + 1}{2} - \frac{\kappa(y_i - 1)}{2} \right) - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\
\text{subject to} \quad & \sum_i \alpha_i y_i = 0 \\
& 0 \leq \alpha_i \leq C C_1; \qquad y_i = 1 \\
& 0 \leq \alpha_i \leq \frac{C}{\kappa}; \qquad y_i = -1
\end{aligned} \tag{51}$$

which reduces to the standard SVM dual when $C_1 = C_{-1} = 1$. Unlike the previous BM-SVM and BP-SVM algorithms, the CS-SVM algorithm performs regardless of the separability of the data and the chosen slack penalty $C$. This can be further studied in detail by writing the dual problem (51) as

$$
\begin{aligned}
\underset{\alpha}{\operatorname{argmax}} \quad & \sum_i \alpha_i^+ + \kappa \sum_i \alpha_i^- - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\
\text{subject to} \quad & \sum_i \alpha_i y_i = 0 \\
& 0 \leq \alpha_i^+ \leq C C_1 \\
& 0 \leq \alpha_i^- \leq \frac{C}{\kappa}
\end{aligned} \tag{52}
$$

with

$$
\begin{aligned}
& 0 < \kappa \leq 1 \leq \frac{1}{\kappa} \leq C_1 \\
& {\alpha_i}^+ = \{\alpha_i | y_i = 1\}, \quad {\alpha_i}^- = \{\alpha_i | y_i = -1\}.
\end{aligned} \tag{53}
$$

Moreover, since $\alpha_i \geq 0$ and $\kappa = 1 - (1 - \kappa)$ we can rewrite (52) with an $\ell_1$-norm norm term as

$$
\begin{aligned}
\underset{\alpha}{\operatorname{argmax}} \quad & -\frac{1}{2} \alpha^T Y K Y \alpha + \mathbf{1}^T \alpha - (1 - \kappa) \|\alpha^-\|_1 \\
\text{subject to} \quad & \alpha^T y = 0 \\
& 0 \preceq \alpha^+ \preceq C C_1 \\
& 0 \preceq \alpha^- \preceq \frac{C}{\kappa}.
\end{aligned} \tag{54}
$$

where $Y = Diag(y)$ and $\mathbf{1}$ is the vector of all ones.

When $C_1 = 1$ and $\kappa = 1$, i.e. $C_{-1} = 1$, the problem of (54) reverts to the standard SVM dual formulation. This implies that (54) is totally compatible with standard dual solvers and its implementation on existing SVM dual solvers is a non-issue.

If we transform problem (54) into a minimization problem, the term $\|\alpha^-\|_1$ acts as an $\ell_1$-norm regularization term with positive coefficient $(1 - \kappa)$. Another difference with the standard cost insensitive SVM (CI-SVM) and BP-SVM dual problem is that in (54), the upper bounds on $\alpha^+$ and $\alpha^-$ are scaled differently. In particular, because $\frac{1}{\kappa} \leq C_1$, the active upper bound constraints on $\alpha_i^+$ are relaxed, compared to $\alpha_i^-$. In summary, the CS-SVM dual problem (54) has two major differences compared to the CI-SVM dual problem:

1. $\ell_1$-norm regularization on $\alpha^-$.

2. relaxed inequality constraints on $\alpha^+$.

These modifications have nontrivial consequences which connect regularization theory and sensitivity analysis to cost-sensitive learning. We study the implications of these modifications by first representing the CI-SVM dual problem as a regularized risk minimization problem which allows us to explain the extra regularization term $-(1-\kappa)\|\alpha^-\|_1$ for both the case of cost sensitive learning and imbalanced learning problems. Subsequently, we study the affect of relaxing the inequality constraint on $\alpha^+$ using sensitivity analysis.

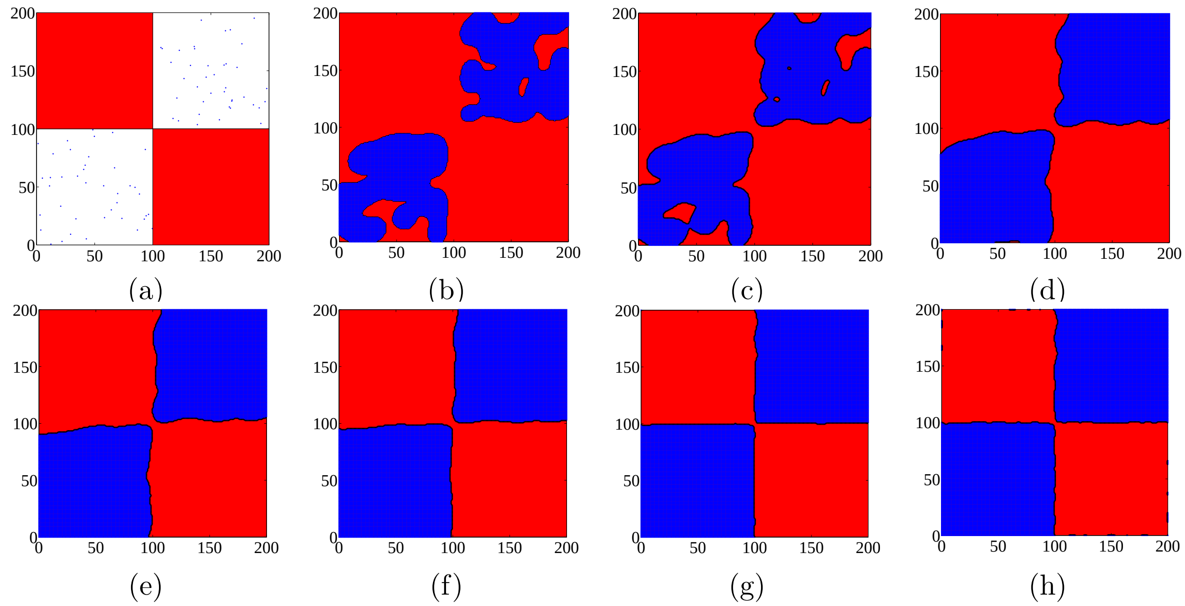


Figure 2: (a) The Checkerboard dataset with imbalance ratio 1:1000, (b) classification result of BM-SVM (c) classification result of BP-SVM with $C_1 = 100$ and $C_{-1} = 1$ (CS-SVM with $\kappa = 1$), (d) classification result of CS-SVM with $\kappa = 0.5$, (e) classification result of CS-SVM with $\kappa = 0.25$, (f) classification result of CS-SVM with $\kappa = 0.1$, (g) classification result of CS-SVM with $\kappa = 0.01$, (h) classification result of CS-SVM with $\kappa = 0.001$.

### 5.1 Regularization on Lagrange multipliers

In this subsection we study the effects of $\ell_1$-norm regularization on $\alpha^-$ in the dual problem, while considering imbalanced dataset learning and cost-sensitive learning separately.

#### 5.1.1 Imbalanced dataset learning

In many applications examples from the target (positive) class are outnumbered by the non-target class. Moreover, in multi-class classification problems where the number of classes are large and a one-versus-all scheme is used, the number of examples in each individual class is usually small compared to the rest of the examples, leading to a highly imbalances problem. These sorts of imbalances occur with different intensity, with ratios between the minority and majority class ranging from 1:10 to 1:$10^6$ Provost and Fawcett (2001).

For the SVM training problem, the number of support vectors grows linearly with the number of examples Steinwart (2004), and this implies that the number of support vectors for each class grows linearly with the number of examples of that class. Therefore, the same imbalance, if not worse, happens in the number of nonzeros in of the solution. In other words, when the dual problem is solved, most of the support vectors belong to the majority class. The problem becomes more apparent when we take into account the equality constraint of (54)

$$\sum_i \alpha_i y_i = 0, \tag{55}$$

which implies

$$\|\alpha^+\|_1 = \|\alpha^-\|_1 \tag{56}$$

Also, results of Steinwart (2004) implies that for imbalanced datasets

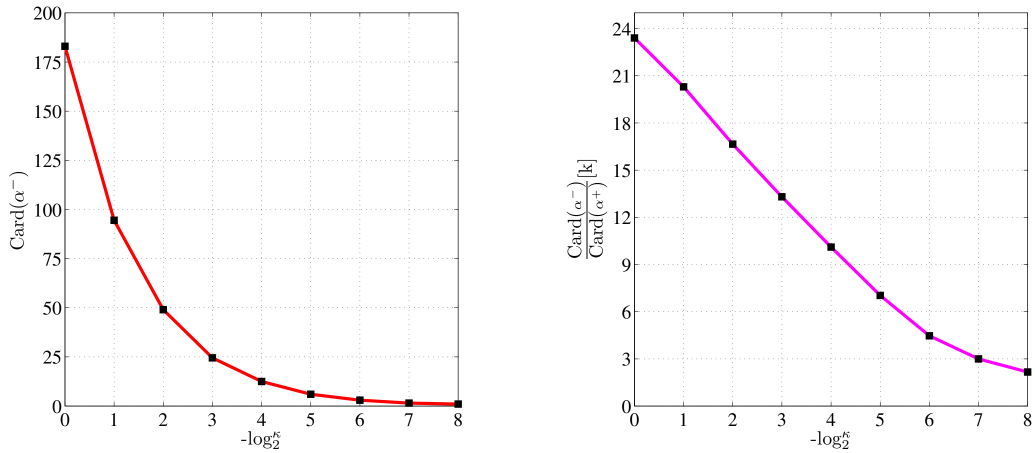
$$\|\alpha^+\|_0 \ll \|\alpha^-\|_0. \tag{57}$$

This results in an irregular solution, with the $\alpha_i^+$s taking values close to the upper bound $C$ and the $\alpha_i^-$s taking values close to the lower bound zero. Wu and Chang (2005) illustrated this problem by conducting an experiment on a 2D Checkerboard dataset with different imbalance ratios as seen in Figure 2(a) . They showed that in the case of imbalanced data, the decision boundary is unwillingly shifted toward the minority class. This is because of a lack of enough examples (support vectors) for the minority class that reside close to the correct decision boundary. When enough examples don't exist at the right place, the margin relies on other examples farther away from the ideal decision boundary, resulting in the decision boundary shifting toward the minority class. They also equivalently illustrated that this is caused by irregular values in the dual variables. This problem persists in the BM-SVM and BP-SVM formulation as a result of their flawed implementation of the asymmetric margin, and can be seen in Figure 2 which show the classification results for the BM-SVM, BP-SVM and CS-SVM on the Checkerboard dataset.

Given that for imbalanced dataset problems the vector $\alpha^-$ has small non sparse elements while the vector $\alpha^+$ is highly sparse (57), the natural remedy is to regularize the non-sparse part of the solution, $\alpha^-$, with a sparsity inducing $\ell_1$-norm regularizer Boyd and Vandenberghe (2004). This leads to a sparse $\alpha^-$, at the solution which is now both balanced and regularized. The CS-SVM problem (54) uses the same technique to deal with the problem of imbalanced datasets by choosing appropriate choice of $\kappa$. As $\kappa$ tends to zero the regularization coefficient $(1-\kappa)$ increases resulting in an increased regularization of the $\alpha_i^-$s, which enforces larger margin for minority (positive) class. Figure 2(g) shows that for a highly imbalanced checkerboard data, an small $\kappa = 0.01$ corrects the decision boundary, close to the optimal one. Choosing $\kappa < 0.01$ violates the condition (53) and has a diminishing return, i.e., leads to preferring the majority class as shown in Figure 2(h).

Also, Figure 3 illustrates the effect of the CS-SVM regularization on the number of support vectors of each class in the solution. The CS-SVM algorithm with different choices of $\kappa$ is applied to the covertype UCI dataset which is imbalanced with a ratio of 1:211, which as the regularization coefficient $(1-\kappa)$ increases, $\alpha^-$ becomes sparser (Figure 3(a)). This leads to an equivalence between the number of non-zero components of $\alpha^-$ and $\alpha^+$ (Figure 3(b)).

Therefore, the CS-SVM in the dual, applies a sparsity inducing $\ell_1$-norm regularization on the $\alpha^-$ and when dealing with imbalanced datasets, the CS-SVM implicitly prevents unwanted movement of the discriminant boundary toward the minority class by enforcing margin to be asymmetric.




(a) (b)

Figure 3: The CS-SVM algorithm for different choices of $\kappa$ is applied to the covertype UCI dataset which is imbalanced with a ratio of 1:211. Starting at $\kappa = 1$, CS-SVM acts as the BP-SVM. (a) shows the reduction in the number of $\alpha^-$ as $\kappa$ decreases and (b) shows the reduction in the imbalance ratio as $\kappa$ decreases. As $\kappa$ decreases the number of negative support vectors is reduced so that by $\kappa = 2^{-256}$ the imbalance ratio between support vectors approaches 1.

#### 5.1.2 Cost-sensitive learning

As shown in the previous section, regularization of any class results in a smaller margin for that class. So, in the cost-sensitive learning setting which costs are known, CS-SVM reduces the margin for the class with the lower cost, or equivalently increases the margin for the class with the higher cost.

In general, the extra $\ell_1$-norm regularization in the CS-SVM dual problem makes the margin asymmetric, in favor of the minority class or the class with higher cost for imbalanced data learning and cost-sensitive learning, respectively.

### 5.2 Regularization on basis expansion coefficients

In the previous section we showed how the Lagrange dual Boyd and Vandenberghe (2004) of the CS-SVM performed $\ell_1$-norm regularization on the support vectors. Rather, in this section we show that the Fenchel dual Rockafellar (1970) of the CS-SVM performs $\ell_1$-norm regularization on the basis coefficients of the discriminant function. A general regularization problem Tikhonov and Arsenin (1977) for given dataset $\mathcal{D}$, loss function $\mathcal{L}$, trade-off hyperparameter $C$, regularizer $\Omega$ and Hilbert space $\mathcal{H}$ can be written as

$$\underset{f\in\mathcal{H}}{\operatorname{argmin}} \;\; \Omega(f) + \mathcal{L}(f;\mathcal{D},C) \tag{58}$$

$$
\begin{tikzcd}[column sep=huge, row sep=huge]
\textbf{Primal}\ \text{Max. Margin Sep.}(w,\Psi,\phi)\text{``?}: kijunds \arrow[r, leftrightarrow, "w^T\Psi_i\leftrightarrow\beta^T K_i", "\Psi^T\Psi\leftrightarrow K"'] \arrow[d, leftrightarrow, "Dual"'] & \text{Reg. Risk Min.}(\beta,K,\phi) \arrow[d, leftrightarrow, "Dual"] \\
\textbf{Dual}\ \text{Dual of: Max. Margin Sep.}(\alpha,K,\phi^*) \arrow[r, leftrightarrow, "\alpha\leftrightarrow YK^{-1}z", "K\leftrightarrow K^{-1}"'] & \text{Reg. Risk Min.}(z,K^{-1},\phi^*)
\end{tikzcd}
$$

Figure 4: The commutative diagram for existing SVM formulations essentially depends on associated parameter spaces $w,\alpha,\beta,z$ and feature spaces $\Psi, K, K^{-1}$. The matrix $\Psi^T$ is the Cholesky factor of $K$, i.e. $K=\Psi^T\Psi$, with its $i^{th}$ row corresponding to the feature space representation of the example $x_i$, i.e., $\Psi_i=\varphi(x_i)$.

which by representer theorem Schölkopf and Smola (2001), (60) has a minimizer of form of

$$f(\mathbf{x})=\sum_{x_i\in\mathcal{D}}\beta_i K(\mathbf{x},x_i)+b. \tag{59}$$

which for Hing loss $\phi$, the primal problem becomes Chapelle (2007):

$$\operatorname*{argmin}_{\beta,b}\ \ \frac{1}{2}\beta^T K\beta+\sum_i\phi(y_i(\beta^T K_i+b)) \tag{60}$$

where $K_i$ is the $i^{th}$ column of the kernel matrix. As shown in the Appendix A, the Fenchel dual problem of (60) can be written as

$$\operatorname*{argmax}_{z}\quad -\Omega^*(g)-\sum_i\phi^*(y_i g_i) \tag{61}$$

which $z\in\mathbb{R}^n$ is dual variable, and $g=K^{-1}z$ is the dual decision function. Figure 4 depicts the relationship between problem (63) for the existing SVM formulations.

As shown in the Appendix A, the CS-SVM dual problem can be written as a regularized risk minimization problem

$$\operatorname*{argmax}_{z}\quad -\Omega^*(g)-\sum_i\phi^*_{\text{BP}}(y_i g_i)-(1-\kappa)\|g^-\|_1 \tag{62}$$

which $g^-$ is a vector of $g_i$s which $y_i=-1$.

Also, by substituting $\Omega^*$, $\phi^*_{\text{BP}}$ (see Appendix A) and setting [1] $g=K^{-1}z$ we have

$$
\begin{aligned}
\operatorname*{argmax}_{z}\qquad & -\frac{1}{2}z^T K^{-1}z+y^T g-(1-\kappa)\|g^-\|_1\\
\text{subject to}\quad & \|g^+\|_1=\|g^-\|_1\\
& 0\preceq g^+\preceq CC_1\\
& 0\preceq -g^-\preceq\frac{C}{\kappa}
\end{aligned} \tag{63}
$$

1. Note that $f(x_i)=K_i^T\beta$ and $g_i=g(x_i)=z^T K_i^{-1}$ are primal and dual decision functions.

There are several points to make:

- By setting $YK^{-1}z = Yg = \alpha$, we can retrieve the SVM's dual problem (54), which by using the fact that $\beta = Y\alpha$ Chapelle (2007), we have $g = \beta$ in (59), (60) and (63). This reveals an interesting duality property: $f = z$ and $g = \beta$, i.e. primal variable is equal to dual decision function and vice versa.

- The $\ell_1$-norm regularization term and equality constraint in (63) can be regarded w.r.t either basis expansion coeeficients $\beta$ or dual decision values $g$.

Compared to CI-SVM and BP-SVM dual problems, problem (63) performs an extra $\ell_1$-norm regularization on the basis expansion coefficients which has a different interpretation in imbalanced data learning and cost-sensitive learning:

**Imbalanced Data Learning** The quantities $\|\beta^+\|_0$ and $\|\beta^-\|_0$ reflect the number of basis functions of each class which contribute the decision function, which similar to their $\alpha$-counter parts they are highly imbalanced, i.e., $\|\beta^+\|_0 \ll \|\beta^-\|_0$. This means that the discriminant function is mostly made up of data-dependent kernel bases of the majority class. which leads to over train on the majority class while under training the minority class. Similar to basis pursuit Chen et al. (1999), CS-SVM adds a $\ell_1$-norm regularization on the basis expansion coefficients to alleviate the problem of over-training on the majority class by balancing the number of basis functions contributing to the decision function.

**Cost-sensitive learning** In the cost-sensitive learning setting, errors of misclassifying one class is higher than the other class and we can translate this to the learning algorithm by choosing more basis functions of the target class. This idea can be implemented by performing $\ell_1$-norm regularization on the expansion coefficients of the lower-cost class ($\beta^-$) (63).

## 6. Example-dependent cost-sensitive learning

In many applications such as computational advertising Agarwal (2011), medical diagnosis Turney (2000), information retrieval Martin Szummer (2011), fraud detection Fawcett and Provost (1997); Stolfo et al. (2000) and business decision-making Zadrozny et al. (2003) the cost of misclassifying an individual example differs from other examples including those of the same class. This gives rise to the concept of example-dependent cost-sensitive (ED-CS) learning.

There main methods to ED-CS learning is *direct cost-sensitive method* Zadrozny and Elkan (2001), which considers a threshold for each example according to its costs, i.e.

$$h(x_i, C_1, C_{-1}) = \begin{cases} 1, & \frac{\eta}{1-\eta} \geq \frac{C_{-1}}{C_1} \\ -1, & \text{otherwise} \end{cases} \tag{64}$$

*MetaCost* Domingos (1999) changes the labels of training set according to (64), and then trains with the new labels. Zadrozny et al. (2003) and Brefeld et al. (2003) proposed methods where the training examples are resampled according to the example cost probability distribution of the data. Despite their simplicity, resampling methods may suffer from over fitting caused by duplicate examples. More recently, Scott (2011) proposed, but did not to implemented, an example-based version of BP-SVM loss function which we call ED-BP-Hinge. The ED-BP-Hinge loss is defined for each example with label $y$, decision value $f$ and cost $c$ as

$$\phi(y, f, c) = c\lfloor 1 - yf \rfloor_+ \tag{65}$$

In dealing with the example dependent cost sensitive learning problem we extend the CS-SVM loss of (45) to the ED-CS-Hinge defined as

$$\phi(y, t, c) = \begin{cases} c\lfloor 1 - yt \rfloor_+, & \text{for } \ y = +1, \\ \lfloor 1 - (2c-1)yt \rfloor_+, & \text{for } \ y = -1. \end{cases} \tag{66}$$

the ED-CS-Hinge loss function inherits the benefits of the CS-SVM loss including the added flexibility of choosing an asymmetric margin of the loss when compared to the ED-BP-Hinge. In the experimental study we implement an example dependent cost sensitive SVM based on the ED-CS-Hinge loss and show an improvement over the ED-BP-Hinge based SVM and other SVM based algorithms on the KDD98 dataset.

## 7. Performance measure

The evaluation of cost sensitive algorithms requires a flexible performance measure that can incorporate different costs and priors. We adopt the cost sensitive zero-one risk which can be written as

$$\begin{aligned} R_{CS} &= E_{Y,\mathbf{X}}[L_{C_1,C_{-1}}(f(\mathbf{x}), y)|\mathbf{X} = \mathbf{x}] \\ &= \sum_y \sum_{\mathbf{x}} P_{\mathbf{X}|\mathbf{Y}}(\mathbf{X} = \mathbf{x}|Y = y) P_{\mathbf{Y}}(y) L_{C_1,C_{-1}}(f(\mathbf{x}), y) \\ &= \sum_y P_{\mathbf{Y}}(+1) \sum_{\mathbf{x}} P_{\mathbf{X}|\mathbf{Y}}(\mathbf{X} = \mathbf{x}|Y = +1) L_{C_1,C_{-1}}(f(\mathbf{x}), +1) \\ &+ \sum_y P_{\mathbf{Y}}(-1) \sum_{\mathbf{x}} P_{\mathbf{X}|\mathbf{Y}}(\mathbf{X} = \mathbf{x}|Y = -1) L_{C_1,C_{-1}}(f(\mathbf{x}), -1) \\ &= P_1 C_1 P_{FN} + P_{-1} C_{-1} P_{FP} \end{aligned} \tag{67}$$

where $P_1$ and $P_{-1}$ are the class priors and $P_{FN}$ and $P_{FP}$ are the false negative and false positive rates respectively. This performance measure readily simplifies to the well known probability of error measure $R_{CI} = P_1 P_{FN} + P_{-1} P_{FP}$, which we call cost insensitive risk.

Finding the best cost sensitive zero-one risk of (67) can be as an instance of vector optimization problem. Each classifier produces a set of vectors $(P_{FP}, P_{FN})$ which should be compared w.r.t. in nonnegative orthant, i.e., $\mathbb{R}^2_+$) which induces component wise inequality in $\mathbb{R}^2_+$. The minimal elements of this set comprise the Pareto optimal frontier Boyd and Vandenberghe (2004) which is also known as the ROC curve in detection theory. Different points on the ROC of a classifier can be found by the vector scalarization optimization problem of

$$\min_{P_{FP}, P_{FN}} \ \lambda_1 P_{FP} + \lambda_2 P_{FN} \tag{68}$$

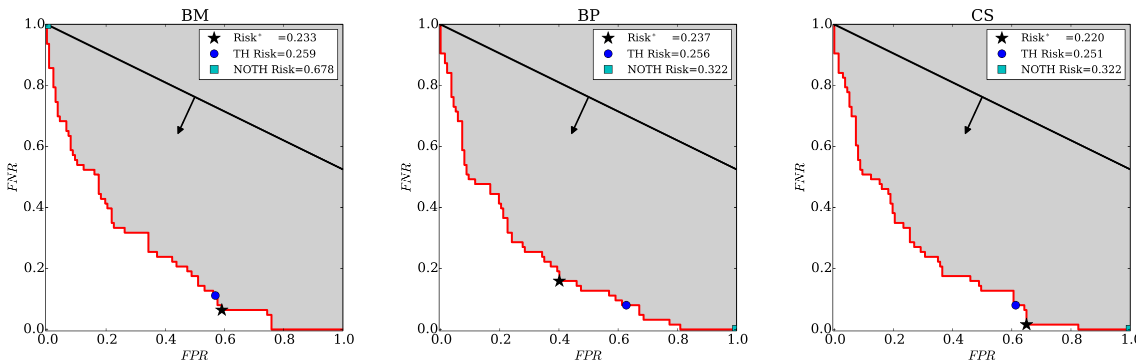



Figure 5: ROC curve of CI-SVM (left), BP-SVM (middle), and CS-SVM (right) on test examples of german dataset which costs are known. For given costs, the objective function is depicted with a black line and the best operating points is shown by Risk*. Also, risks associated with the models which tuned by a threshold in the training phase is denoted by TH-Risk and risk of the models without thresholding is also shown (NOTH-Risk).

Choosing $(\lambda_1, \lambda_2) = (P_1 C_1, P_{-1} C_{-1})$ results in the following optimization problem

$$\min_{P_{FP}, P_{FN}} \quad P_1 C_1 P_{FP} + P_{-1} C_{-1} P_{FN}. \tag{69}$$

which has an objective function equal to the cost sensitive zero-one risk of (67). This means that by using the cost sensitive zero-one risk as the performance measure and choosing a certain $(P_1 C_1, P_{-1} C_{-1})$ we are in fact finding a certain optimal point on the classifier ROC curve that corresponds to $(\lambda_1, \lambda_2) = (P_1 C_1, P_{-1} C_{-1})$. We use the term minimum risk instead of minimum cost-sensitive zero-one risk in the rest of the paper.

When the $(P_1 C_1, P_{-1} C_{-1})$ are known, we simply use them in the evaluation of the classifier as well as finding the best threshold Figure 7.

When the costs or priors of a problem are not known, a single point on ROC curve might not be a robust performance measure for the classifier. So we evaluate the risk at all points within a low $FP$ or low $TP$of the ROC. This is equivalent to finding the $t$-AUC Wu et al. (2008) which evaluates the area under the ROC curve within the 1 to $t$ true negative regions. we extend this method and propose the TP-$t$-AUC and TN-$t$-AUC to evaluate the area under the ROC curve within the 1 to $t$ true positive and 1 to $t$ true negative regions respectively. In the experiments we specifically report both the TP-$t$-AUC and TN-$t$-AUC in order to demonstrate the CS-SVM's ability in learning models with both high sensitivity and high specificity.

## 8. Experimental study

In this section we conduct extensive experiments on 21 real world datasets and compare the BM-SVM, BP-SVM and CS-SVM algorithms. The experiments are grouped into four types namely cost-sensitive learning with available class-dependent costs (CSA), cost-sensitive learning when class-dependent costs are unavailable(CSU), cost-sensitive learning with example-dependent costs (CSE) and imbalanced dataset learning(IDL). The datasets and experiments are further explained in the following sections.

### 8.1 Datasets

21 datasets, created from 20 distinct datasets, are used to compare the performance of the CS-SVM algorithm with other algorithms under different scenarios. Table 1 shows the detailed specifications of each dataset. Each dataset is associated with a type of experiment. For example, the KDD98 dataset is used in the CSE experiment and datasets with large class imbalance ratios are used in IDL experiments. For each dataset we choose the class with the higher cost or fewer data points as the target or positive class. All multi-class datasets were converted to binary datasets. In particular, the binary datasets SIAM(1) and SIAM(2) are datasets which have been constructed from the same multi-class dataset but with different target class and thus different imbalance ratios. [2]

### 8.2 Setup

The RBF Gaussian kernel $k(x, x') = \exp -\gamma\|x - x'\|^2$ is used for all SVM algorithms. We choose the hyperparameters of $C$ and $\gamma$ by performing a 2D grid search and optimizing the associated performance measure (risk, TP/TN-$t$-AUC or income). A more elaborate algorithm for hyperparameter selection in CS-SVMs is developed by Gu et al. (2015, 2017). Given that the size of the datasets are very different, we avoid over fitting by considering a specific search range and granularity for each dataset, but use the same range and granularity for all algorithms. In each iteration of the grid search, the performance is evaluated by 10 fold cross-validation for small datasets and evaluated on a separated validation set for large datasets which appear in bold font in Table 1. Once the 2D grid search is complete, the hyper parameters are used to train the BM-SVM. Also, the kernel hyper parameter is used for training both the BP-SVM and the CS-SVM.

Without loss of generality, we set $C_{-1} = 1$ in the BP-SVM experiments. Therefore, when considering the BP-SVM experiments we only need to perform an additional 2D grid search for $C$ and $C_1$. The CS-SVM actually has four independent hyper parameters, including $\gamma$. We perform a 3D grid search on $C$, $C_1$ and $\kappa$ when the costs are not known, and a 2D search on $C$ and $\kappa$ when the costs $C_1$ and $C_{-1}$ are available. Note that in the case of available costs (CSA), setting $\kappa$ to a value other than $\kappa = \frac{1}{2C_{-1}-1}$ implicitly means that $C_{-1}$ is set to a value other than its determined value. However, we deliberately allow this in order to make use of the CS-SVM algorithm's asymmetric margin advantages. Nevertheless, we use the determined cost of $C_{-1}$ during performance evaluation. [3]. Finally, we use the TP-0.9-AUC and TN-0.9-AUC performance measures when considering the IDL and CSU type experiments since the costs are not explicitly known in these experiments.

2. SIAM, Web Spam, IJCNN, MNIST, KDD99 and Covertype data sets were obtained from the LIBSVM data website.http://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets
3. The source code for CS-SVM is available at http://www.svcl.ucsd.edu/projects/costlearning

Table 1: Specifications of the benchmark datasets. # of Ex. is the number of example data points. # of Feat. is the number of features. Ratio is the class imbalance ratio. Target specifies the target or positive class. Type specifies the type of experiment conducted on this dataset.

| Dataset | # of Ex. | # of Feat. | Ratio | Target | Type |
|---|---|---|---|---|---|
| German Credit | 1,000 | 24 | 1:2 | Bad (2) | CSA |
| Heart | 270 | 13 | 1:1 | Presence (2) | CSA |
| **KDD 99 (Intrusion Detection)** | 5,209,460 | 42 | 1:4 | Normal | CSA |
| **KDD 98 (Donation)** | 191,779 | 479 | 1:20 | 2 | CSE |
| Breast Cancer Diagnostic | 569 | 32 | 1:2 | Malignant (M) | CSU |
| Breast Cancer Original | 699 | 10 | 1:2 | Malignant (4) | CSU |
| Diabetes | 768 | 8 | 1:2 | Has Diabet (+1) | CSU |
| Echo-cardiogram | 132 | 12 | 1:2 | Alive (1) | CSU |
| Liver | 345 | 6 | 1:1 | 1 | CSU |
| Sonar | 208 | 60 | 1:1 | +1 | CSU |
| Tic-Tac-Toe | 958 | 9 | 1:2 | Negative | CSU |
| **Web Spam** | 350,000 | 254 | 1:2 | -1 | CSU |
| Breast Cancer Prognostic | 198 | 34 | 1:3 | Recur ( R ) | IDL |
| **Covertype** | 581,012 | 54 | 1:211 | Cottonwood/Willow(4) | IDL |
| Hepatits | 155 | 20 | 1:4 | Die (1) | IDL |
| **IJCNN** | 141,691 | 2 | 1:10 | +1 | IDL |
| **Isolet** | 7,797 | 617 | 1:25 | K (11) | IDL |
| **MNIST** | 70,000 | 780 | 1:10 | 5 | IDL |
| **SIAM1** | 28,596 | 30438 | 1:2000 | 1,6,7,11 | IDL |
| **SIAM11** | 28,596 | 30438 | 1:716 | 11,12 | IDL |
| Survival | 306 | 3 | 1:3 | 2 | IDL |

### 8.3 Implementation

The CS-SVM problem (54) is readily implemented in the dual by modifying the LibSVM Chang and Lin (2011) source code. This is done by 1) adding the regularization term to the LibSVM objective function and 2) selecting $C_1 = C_1$ and $C_{-1} = \frac{1}{\kappa}$ as the cost parameters. As a result, $C$, $\gamma$, $C_1$ and $\frac{1}{\kappa}$ are the CS-SVM solver hyper parameters.

### 8.4 Experiments on cost-sensitive learning with known class-dependent costs

For these set of experiments, we compare test Risk of datasets corresponding to the point on ROC curve which determined by the threshold that is found in the training phase (Figure 7). Three datasets with known class costs are examined. Namely, the German credit card dataset Geibel et al. (2004); Newman et al. (1998), the Statlog Heart Disease Newman et al. (1998) and KDD99 Elkan (2000) datasets are considered. The minimum risk using the BM-SVM, BP-SVM and CS-SVM is shown in Table 2 for each of the CSA datasets. The CS-SVM algorithm outperforms the BP-SVM on all datasets, surpasses the BM-SVM on two and ties with the BM-SVM on one dataset.

Table 2: Expected risk of datasets with known class-dependent costs.

| Dataset | BM-SVM | BP-SVM | CS-SVM | SVM+SMOTE | CS-SVM+SMOTE |
|---|---|---|---|---|---|
| German Credit | 0.26 | 0.6 | **0.25** | | |
| Heart | **0.09** | 0.1 | **0.09** | | |
| KDD 99 | 0.054 | 0.054 | **0.045** | | |

Table 3: TP-0.9-AUC on datasets with unknown class costs.

| Dataset | BM-SVM | BP-SVM | CS-SVM | SVM+SMOTE | CS-SVM+SMOTE |
|---|---|---|---|---|---|
| Breast Cancer D. | 0.988 | 0.989 | **0.991** | 0.984 | 0.988 |
| Breast Cancer O. | 0.993 | **0.994** | **0.994** | 0.984 | 0.992 |
| Breast Cancer P. | 0.659 | 0.669 | **0.686** | 0.610 | 0.666 |
| Covertype | 0.976 | **0.993** | **0.993** | 0.855 | 0.992 |
| Diabetes | 0.858 | 0.881 | 0.887 | 0.848 | **0.891** |
| Echo-cardiogram | 0.938 | 0.950 | **0.951** | 0.770 | 0.933 |
| Hepatitis | 0.760 | 0.787 | **0.907** | 0.873 | 0.887 |
| IJCNN | 0.989 | 0.991 | 0.993 | **0.994** | **0.994** |
| ISOLET | 0.995 | **0.997** | **0.997** | 0.976 | 0.976 |
| Liver | 0.778 | 0.784 | 0.787 | 0.773 | **0.791** |
| MNIST | 0.992 | **0.994** | **0.994** | 0.990 | 0.992 |
| SIAM1 | 0.718 | 0.743 | **0.752** | 0.671 | 0.733 |
| SIAM11 | 0.680 | 0.698 | **0.718** | 0.669 | 0.685 |
| Sonar | 0.917 | 0.917 | 0.917 | 0.917 | **0.938** |
| Survival | 0.680 | 0.697 | **0.767** | 0.713 | 0.755 |
| Tic-Tac-Toe | 0.903 | **1.000** | **1.000** | **1.000** | **1.000** |
| Web Spam | 0.989 | 0.989 | **0.991** | 0.984 | **0.991** |

### 8.5 Experiments on cost-sensitive learning with unknown class-dependent costs

We consider eight datasets which do not have known costs and are not highly imbalanced. Namely, we examine the Breast Cancer Diagnostic, Breast Cancer Original, Pima Indian Diabets, Echo-cardiogram, Liver, Sonar, Tic-Tac-Toe Newman et al. (1998) and Web Spam Webb et al. (2006) datasets. The CS-SVM exhibits improved TP-0.9-AUC (Table 3) and TN-0.9-AUC (Table 4) performance compared to BP-SVM and BM-SVM in 15 out of 16 experiments and ties in one experiment.

### 8.6 Experiments on imbalanced data learning

We examine large datasets with severe imbalance ratios to evaluate the merit of the proposed CS-SVM algorithm on imbalanced data learning which could be the most prevailing problem in practice. The CS-SVM exhibits improved TP-0.9-AUC (Table 3) and TN-0.9-AUC (Table 4) performance compared to BP-SVM and BM-SVM in 17 out of 18 IDL experiments and ties in one experiment.

Table 4: TN-0.9-AUC on datasets with unknown class costs.

| Dataset | BM-SVM | BP-SVM | CS-SVM |
|---|---|---|---|
| Breast Cancer D. | 0.40 | 0.35 | **0.31** |
| Breast Cancer O. | 0.17 | 0.17 | **0.16** |
| Diabetes | 0.69 | 0.67 | **0.66** |
| Echo-cardiogram | 0.60 | 0.60 | **0.35** |
| Liver | 0.90 | 0.95 | **0.88** |
| Sonar | 0.70 | 0.62 | **0.60** |
| Tic-Tac-Toe | 0.93 | 0.87 | **0.86** |
| Web Spam | 0.03 | 0.03 | **0.02** |

Table 5: TP-0.9-AUC on imbalanced datasets.

| Dataset | BM-SVM | BP-SVM | CS-SVM | |
|---|---|---|---|---|
| Breast Cancer P. | 0.83 | 0.79 | **0.76** | IDL |
| Covertype | 0.034 | 0.020 | **0.016** | IDL |
| Hepatits | 0.56 | 0.40 | **0.36** | IDL |
| IJCNN | 0.091 | 0.034 | **0.031** | IDL |
| Isolet | 0.86 | 0.19 | **0.10** | IDL |
| MNIST | 0.053 | 0.019 | **0.017** | IDL |
| SIAM1 | 0.76 | 0.30 | **0.29** | IDL |
| SIAM11 | **0.70** | **0.70** | **0.70** | IDL |
| Survival | 0.89 | 0.88 | **0.87** | IDL |

### 8.7 Experiments on cost-sensitive learning with example-dependent cost

We study example-dependent cost-sensitive learning using the well known KDD98 dataset. This dataset contains information about past contributors to charities. The task is to classify individuals as donors or non-donors for a new charity so that overall donations are maximized. The cost of sending mail and soliciting a donation is 0.68$ and the range of possible donations is $1-200$$. We use the total profit performance measure Elkan (2001) and evaluate the algorithms according to the benefit matrix shown in Table 7.

A range of different methods and algorithms have been previously used on this dataset and some of the most profitable methods are listed in Table 8 and further explained. Wong et al. (2005) proposed an ad-hoc algorithm which extracts Focused Association Rules (FAR) for the KDD98 dataset. The FAR method consist of three subsequent algorithms of rule generating, model building and pruning and yields the best profit on the KDD98 dataset. The example dependent MetaCost (ED-MetaCost) and direct cost-sensitive method (DCSM) are both implemented by Zadrozny and Elkan (2001) and differ in the method used for cost and probability estimation. Res-DIPOL and Res-ED-BP-SVM Geibel et al. (2004) are resampling based algorithms equipped with DIPOL and ED-BP-SVM algorithms respectively. For these methods the dataset is resampled according to a modified probability distribution. Zadrozny et al. (2003) suggest two types of algorithms for cost sensitive learning. The first type are those that directly incorporate the costs into the learning algorithm and the second type are black box methods that convert a cost insensitive algorithm into a cost sensitive algorithm by resampling the data according to the example costs. The

Table 6: TN-0.9-AUC on imbalanced datasets.

| Dataset | BM-SVM | BP-SVM | CS-SVM | |
|---|---|---|---|---|
| Breast Cancer P. | 0.87 | 0.81 | **0.80** | IDL |
| Covertype | 0.062 | 0.062 | **0.060** | IDL |
| Hepatits | 0.70 | 0.70 | **0.67** | IDL |
| IJCNN | 0.02 | 0.02 | **0.01** | IDL |
| Isolet | 0.86 | 0.19 | **0.10** | IDL |
| MNIST | 0.05 | 0.02 | **0.02** | IDL |
| SIAM1 | 0.938 | 0.526 | **0.525** | IDL |
| SIAM11 | 1.000 | 0.748 | **0.739** | IDL |
| Survival | 0.66 | 0.64 | **0.63** | IDL |

Table 7: Benefit matrix for the KDD98 dataset.

| | Donor | Non-donor |
|---|---|---|
| Predicted Donor | $C_{+1_i}\$$ | $-0.68\$$ |
| Predicted Non-donor | $-C_{+1_i}\$$ | $0\$$ |

Polynomial kernel ED-BP-SVM (P-ED-BP-SVM) directly incorporates the costs into the learning algorithm while the proposed black box SVM (BB-CI-SVM) and black box C4.5 (BB-C4.5) are examples of the second type proposed in Zadrozny et al. (2003).

Table 7 also shows results for the example dependent implementations of BM-SVM (ED-BM-SVM), BP-SVM (ED-BP-SVM) and CS-SVM (ED-SV-SVM) with Gaussian kernels. The ED-CS-SVM exhibits the best performance among all ED-SVM methods. It also ranks fifth among all methods some of which use complicated and compounded schemes.

## 9. 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 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. The dual problem of CS-SVM is studied and connections between cost-sensitive learning and regularization theory and sensitivity analysis are established. Minimum expected cost-sensitive risk is considered as a metric for evaluating the performance of binary classifiers in the cost-sensitive and imbalanced data settings. The CS-SVM is also readily extended to cost-sensitive learning with example-dependent costs. Empirical evidence confirms its superior performance, when compared to previous methods.

Table 8: Income of different algorithms on the KDD98 dataset.

| Rank | Algorithm | Income | Comments |
|---|---|---|---|
| 1 | FAR | \$ 20,693 | Ad-hoc method based on sequence of three algorithms |
| 2 | DCSM | \$ 15,329 | Probability and cost estimation to minimize cost |
| 3 | BB-C4.5 | \$ 15,016 | C4.5 on resampled dataset |
| 4 | KDD-Cup 98 Winner | \$ 14,712 | Rule-based approach |
| 5 | ED-CS-SVM | \$14,205 | ED-CS-SVM with Gaussian kernel $\kappa = 0.97$ |
| 6 | ED-MetaCost | \$ 14,113 | Probability and cost estimation to minimize cost |
| 7 | ED-BP-SVM | \$14,008 | ED-BP-SVM with Gaussian kernel |
| 8 | Res-DIPOL | \$ 14,045 | DIPOL on resampled dataset |
| 9 | P-ED-BP-SVM | \$ 13,683 | ED-BP-SVM with Polynomial Kernel |
| 10 | BB-SVM | \$ 13,152 | CI-SVM on resampled dataset |
| 11 | Res-ED-BP-SVM | \$ 12,883 | ED-BP-SVM on resampled dataset |
| 12 | BM-CI-SVM | \$ 10,560 | Standard SVM |
| 13 | Null Classifier | \$ 10,560 | Predicts all examples as donor |

## Appendix A. Fenchel Dual Problem

**Theorem 3 (Fenchel Dual of the Regularized risk Minimization Problem)** *Let $\Omega : \mathbb{R}^n \to \mathbb{R}$ and $\phi : \mathbb{R} \to \mathbb{R}_+$ be convex functions and $\mathbf{Dom}\,\Omega = \mathbf{Dom}\,\Phi = \mathbb{R}^n$, then*

$$\inf_{\beta}\{\Omega(K\beta) + \sum_i \phi(y_i K_i^T \beta)\} = \sup_{z}\{-\Omega^*(K^{-1}z) - \sum_i \phi^*(y_i z^T K_i^{-1})\} \tag{70}$$

*which $\beta$ and $z$ are primal and dual variables, and $\Omega^*$ and $\phi^*$ are Fenchel Conjugate functions*[4] *of $\Omega$ and $\Phi$,respectively.*

**Proof**

(i) By the representer therem we have $f(.) = K\beta$.

(ii) Fenchel Duality Theorem Rockafellar (1970) and induction we have

$$\inf_{f}\{\Omega(f) + \sum_i \phi_i(f)\} = \sup_{g}\{-\Omega^*(g) - \sum_i \phi_i^*(g)\} \tag{71}$$

where here $f$ and $g$ are primal and dual decision functions.

(iii) Let $\Phi : \mathbb{R}^n \to \mathbb{R}^n$, then we can write

$$\sum_i \phi(y_i K_i^T \beta) = \mathbf{1}^T \Phi(YK\beta) \tag{72}$$

4. The Fenchel conjugate of $h : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$h^*(y) = \sup_{x \in \mathbf{Dom}\,h} \{y^T x - h(x)\}$$

(iv) Composition with linear transformation can be conjugated by [5] Boyd and Vandenberghe (2004)

$$\begin{aligned}\Omega(f) = \Omega(K\beta) \Rightarrow \Omega^*(g) = \Omega^*(K^{-1}z)\\ \Phi(f) = \Phi(YK\beta) \Rightarrow \Phi^*(g) = \Phi^*(YK^{-1}z)\end{aligned} \tag{73}$$

where $f$ and $g$ are primal and dual decision functions and $\beta$ and $z$ are primal and dual variables, respectively.

(v) By (73), we have $g(.) = K^{-1}z$

(vi) From (71) and (72) we have

$$\mathbf{1}^T\Phi(YK^{-1}z) = \sum_i \phi^*(y_i z^T K_i^{-1}) \tag{74}$$

■

Conjugate of regularizer [6] of $\Omega(K\beta) = \frac{1}{2}\beta^T K\beta$ is given by

$$\Omega^*(K^{-1}z) = \sup_\beta\{z^T K^{-1}K\beta - \frac{1}{2}\beta^T K\beta\} = \frac{1}{2}z^T K^{-1}z$$

For the decision functions with a bias term, i.e. $f(x_i) = K_i^T\beta + b$, the bias is not regularized, and unregularized bias formulation introduces an equality constraint in the dual (Rifkin and Lippert (2007), Section 9.1)

$$\mathbf{1}^T K^{-1}z = 0 \tag{75}$$

Given $a, b \in \mathbb{R}_{++}$, the conjugate of the Hinge loss $\phi(\mathbf{u}) = \max(b - a\mathbf{u}, 0)$, can be computed

$$g^*(\mathbf{v}) = \sup_{\mathbf{u}}\{\mathbf{uv} - \max(b - ax, 0)\} = \begin{cases}\sup\limits_{\mathbf{u}}\ \{\mathbf{uv}\} & \mathbf{u} > \frac{b}{a}\\ \sup\limits_{\mathbf{u}}\ \{\mathbf{u}(\mathbf{v}+a) - b\} & \mathbf{u} \le \frac{b}{a}\end{cases}$$

which we have two cases

$$\text{(i)}\ \ \mathbf{v} \le 0 \ \Rightarrow\ g^*(\mathbf{v}) = \begin{cases}\sup\limits_{\mathbf{u}>\frac{b}{a}}\ \{\mathbf{uv}\} = \frac{b}{a}\mathbf{v}\\ \\ \sup\limits_{\mathbf{u}\le 1}\ \{\mathbf{u}(\mathbf{v}+a) - b\} = \begin{cases}\frac{b}{a}\mathbf{v} & -a \le \mathbf{v} \le 0\\ \infty & \mathbf{v} < -a\end{cases}\end{cases}$$

$$\text{(ii)}\ \ \mathbf{v} > 0 \ \Rightarrow\ g^*(\mathbf{v}) = \begin{cases}\sup\limits_{\mathbf{u}>\frac{b}{a}}\ \{\mathbf{uv}\} = \infty\\ \sup\limits_{\mathbf{u}\le\frac{b}{a}}\ \{\mathbf{u}(\mathbf{v}+a) - b\} = \frac{b}{a}\mathbf{v}\end{cases}$$

Thus for all $\mathbf{u}$, we can write $g^*(\mathbf{v}) = I_{[-1,0]}(\mathbf{v}) + \mathbf{v}$ or equivalently [7] $g^*(\mathbf{v}) = I_{[0,1]}(\mathbf{v}) - \mathbf{v}$. Now we derive the conjugate of CI-Hinge, BP-Hinge and CS-Hinge losses specifically:

5. Note that, $Y$ is a diagonal matrix with $y_{ii} \in \{-1, 1\}$. So we have $Y = Y^{-1}$.
6. $\Phi$ and $\Phi^*$ are both Tikhonov regularization in $\mathcal{H}$ and $\mathcal{H}^*$ with kernels $K$ and $K^{-1}$ respectively, i.e., $\Omega(f) = \frac{1}{2}\|f\|^2_{\mathcal{H}} = \frac{1}{2}f^T K^{-1} f$ and $\Omega^*(g) = \|g\|^2_{\mathcal{H}^*} = \frac{1}{2}z^T K^{-1}z$
7. This equivalence is legitimate because $\Omega$ is an even function.

**CI-Hinge**

$$\phi(\mathbf{u}) = C\max(1-\mathbf{u},0) = \max(C-Cu,0) \;\Rightarrow\; \phi^*(\mathbf{v}) = I_{[0,C]}(\mathbf{v}) - \mathbf{v}$$

**BP-Hinge**

$$\phi_{\text{BP}+}(\mathbf{u}) = CC_1\max(1-\mathbf{u},0) \;\Rightarrow\; \phi^*_{\text{BP}+}(\mathbf{v}) = I_{[0,CC_1]}(\mathbf{v}) - \mathbf{v}$$
$$\phi_{\text{BP}-}(\mathbf{u}) = CC_{-1}\max(1-\mathbf{u},0) \;\Rightarrow\; \phi^*_{\text{BP}-}(\mathbf{v}) = I_{[0,CC_{-1}]}(\mathbf{v}) - \mathbf{v}$$

**CS-Hinge**

$$\phi_{\text{CS}+}(\mathbf{u}) = CC_1\max(1-\mathbf{u},0) \;\Rightarrow\; \phi^*_{\text{CS}+}(\mathbf{v}) = I_{[0,CC_1]}(\mathbf{v}) - \mathbf{v}$$
$$\phi_{\text{CS}-}(\mathbf{u}) = C\max(1-\frac{\mathbf{u}}{\kappa},0) \;\Rightarrow\; {\phi_{\text{CS}}}^*_+(\mathbf{v}) = I_{[0,\frac{C}{\kappa}]}(\mathbf{v}) - \kappa\mathbf{v}$$

Moreover, since $\mathbf{v} \geq 0$, for $C_{-1} = \frac{1}{\kappa}$we can write

$$\phi^*_{\text{CS}}(\mathbf{v}) = \phi^*_{\text{BP}}(\mathbf{v}) + (1-\kappa)|\mathbf{v}|$$

and in general we have

$$\sum_i \phi^*_{\text{CS}}(y_i z^T K_i^{-1}) = \sum_i \phi^*_{\text{BP}}(y_i z^T K_i^{-1}) + (1-\kappa)\|K^{-1}z^-\|_1$$

## References


Deepak Agarwal. Cost-Sensitive Learning in Computational Advertising. In *Cost-Sensitive Machine Learning*, Chapman & Hall/CRC Machine Learning & Pattern Recognition, pages 241—-263+. CRC Press, dec 2011. ISBN 978-1-4398-3925-6. doi: 10.1201/b11423-11.

Rehan Akbani, Stephen Kwek, and Nathalie Japkowicz. Applying Support Vector Machines to Imbalanced Datasets. In *European Conference on Machine Learning (ECML)*, pages 39–50, 2004.

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

Francis R Bach, David Heckerman, and Eric Horvitz. Considering Cost Asymmetry in Learning Classifiers. *The Journal of Machine Learning Research*, 7:1713–1741, 2006.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, Classification, and Risk Bounds. *Journal of the American Statistical Association*, 101(473):138–156, mar 2006. ISSN 0162-1459. doi: 10.1198/016214505000000907.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, mar 2004. ISBN 0521833787.

Ulf Brefeld, Peter Geibel, and Fritz Wysotzki. Support Vector Machines with Example Dependent Costs. In *Proceedings of the European Conference on Machine Learning*, 2003.

Chih-chung Chang and Chih-jen Lin. {LIBSVM}: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:1–39, 2011.

Olivier Chapelle. Training a support vector machine in the primal. In Léon Bottou, Olivier Chapelle, Dennis DeCoste, and Jason Weston, editors, *Large Scale Kernel Machines*, volume 19, book part (with own title) 2, pages 29–50. MIT Press, may 2007. doi: 10.1162/neco.2007.19.5.1155.

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

Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic Decomposition by Basis Pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1999.

Corinna Cortes and Vladimir N Vapnik. Support-Vector Networks. *Machine Learning*, 20:273–297, 1995.

M A A Davenport, R G G Baraniuk, and Clayton D Scott. Controlling False Alarms With Support Vector Machines. In *ICASSP*, 2006.

Pedro Domingos. MetaCost: a general method for making classifiers cost-sensitive. In *Knowledge Discovery and Data Mining*, pages 155–164, 1999.

Charles Elkan. Results of the KDD'99 classifier learning. *ACM SIGKDD Explorations Newsletter*, 1(2):63–64, jan 2000. ISSN 1931-0145. doi: 10.1145/846183.846199.

Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the 17th international joint conference on Artificial intelligence - Volume 2*, pages 973–978, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-812-5, 978-1-558-60812-2.

Seyda Ertekin, Jian Huang, Leon Bottou, and Lee Giles. Learning on the Border: Active Learning in Imbalanced Data Classification. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 127–136, Lisbon, Portugal, 2007. ACM. ISBN 978-1-59593-803-9. doi: 10.1145/1321440.1321461.

Tom Fawcett and Foster Provost. Adaptive Fraud Detection. *Data Mining and Knowledge Discovery*, 1(3):291–316, jan 1997. ISSN 1384-5810. doi: 10.1023/A:1009700419189.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive Logistic Regression: a Statistical View of Boosting. *Annals of Statistics*, 28(2):337–374, 2000.

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

Bin Gu, Victor S Sheng, and Shuo Li. Bi-Parameter Space Partition for Cost-Sensitive SVM. In *IJCAI*, pages 3532–3539, 2015.

Bin Gu, Victor S Sheng, Keng Yeow Tay, Walter Romano, and Shuo Li. Cross Validation Through Two-Dimensional Solution Surface for Cost-Sensitive SVM. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1103–1121, 2017.

Grigoris Karakoulas and John Shawe-Taylor. Optimizing classifiers for imbalanced training sets. In *Proceedings of the 1998 conference on Advances in neural information processing systems II*, pages 253–259, Cambridge, MA, USA, 1999. MIT Press.

Suzan Köknar-Tezel and Longin Jan Latecki. Improving SVM Classification on Imbalanced Data Sets in Distance Spaces. *2009 Ninth IEEE International Conference on Data Mining*, pages 259–267, dec 2009. doi: 10.1109/ICDM.2009.59. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5360251.

Miroslav Kubat and Stan Matwin. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In *ICML*, pages 179–186, 1997.

Yi Lin, Yoonkyung Lee, and Grace Wahba. Support Vector Machines for Classification in Nonstandard Situations. *Machine Learning*, 46:191–202, 2002.

Filip Radlinski Martin Szummer. Cost-sensitive Machine Learning for Information Retrieval. In *Cost-Sensitive Machine Learning*, Chapman & Hall/CRC Machine Learning & Pattern Recognition. CRC Press, dec 2011. ISBN 978-1-4398-3925-6. doi: 10.1201/b11423-11.

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

Hamed Masnadi-Shirazi and Nuno Vasconcelos. On the Design of Loss Functions for Classification: theory, robustness to outliers, and SavageBoost. In *NIPS*, pages 1049–1056, 2008.

Josey Mathew, Chee Khiang Pang, Ming Luo, and Weng Hoe Leong. Classification of Imbalanced Data by Oversampling in Kernel Space of Support Vector Machines. *IEEE transactions on neural networks and learning systems*, 2017.

Javier M Moguerza and Alberto Munoz. Support Vector Machines with Applications. *Statistical Science*, 21:322–336, 2006.

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

J Platt. Probabilistic Outputs for Support Vector Machines and Comparison to Regularized Likelihood Methods. In *Adv. in Large Margin Classifiers*, 2000.

Foster Provost and Tom Fawcett. Robust Classification for Imprecise Environments. *Machine Learning*, 42:203–231, 2001. doi: 10.1023/A:1007601015854.

Ryan M Rifkin and Ross A Lippert. Value Regularization and Fenchel Duality. *Journal of Machine Learning Research*, 8:441–479, 2007. ISSN 1532-4435.

R Tyrrell Rockafellar. *Convex Analysis.* Princeton University Press, 1970. ISBN 0691080690.

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

Bernhard Schölkopf and Alexander J Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, MA, USA, 2001. ISBN 0262194759.

Clayton D Scott. Surrogate losses and regret bounds for cost-sensitive classification with example-dependent costs. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 153–160, Bellevue, Washington, USA, jun 2011. ACM.

Ingo Steinwart. Sparseness of Support Vector Machines—Some Asymptotically Sharp Bounds. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

Salvatore J Stolfo, Wei Fan, Wenke Lee, Andreas Prodromidis, and Philip K Chan. Cost-based Modeling for Fraud and Intrusion Detection: Results from the JAM Project. In *In Proceedings of the 2000 DARPA Information Survivability Conference and Exposition*, pages 130–144, 2000.

Yuchun Tang, Yan-Qing Zhang, Nitesh V Chawla, and Sven Krasser. SVMs modeling for highly imbalanced classification. *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, 39(1):281–288, feb 2009. ISSN 1941-0492. doi: 10.1109/TSMCB.2008.2002909. URL http://www.ncbi.nlm.nih.gov/pubmed/19068445.

Andrei N Tikhonov and Vasilii Y Arsenin. *Solutions of Ill-Posed Problems*. V. H. Winston & Sons, Washington, D.C.: John Wiley & Sons, New York,, 1977.

Peter Turney. Types of Cost in Inductive Concept Learning. In *Proceedings of the ICML'2000 Workshop on Cost-Sensitive Learning*, pages 15–21, 2000.

Vladimir N Vapnik. *Statistical Learning Theory*. John Wiley Sons Inc, 1998.

Senzhang Wang, Zhoujun Li, Wenhan Chao, and Qinghua Cao. Applying adaptive over-sampling technique based on data density and cost-sensitive SVM to imbalanced learning. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8. IEEE, 2012a.

Senzhang Wang, Zhoujun Li, and Xiaoming Zhang. Bootstrap sampling based data cleaning and maximum entropy SVMs for large datasets. In *Tools with Artificial Intelligence (ICTAI), 2012 IEEE 24th International Conference on*, volume 1, pages 1151–1156. IEEE, 2012b.

Steve Webb, James Caverlee, and Calton Pu. Introducing the webb spam corpus: Using email spam to identify web spam automatically. In *Proceedings of the Third Conference on Email and Anti-Spam (CEAS)*, 2006.

Ke Wang K W Wong, Senqiang Zhou, Qiang Yang, and Jack Man Shun Yeung. Mining Customer Value: From Association Rules to Direct Marketing. *Data Mining and Knowledge Discovery*, 11(1):57–79, jul 2005. ISSN 1384-5810. doi: 10.1007/s10618-005-1355-x.

Gang Wu and Edward Y Chang. Adaptive Feature-Space Conformal Transformation for Imbalanced Data Learning. In *Proceedings of the Twentieth International Conference on Machine Learning*, volume 20, pages 816–823, 2003.

Gang Wu and Edward Y Chang. KBA: Kernel Boundary Alignment Considering Imbalanced Data Distribution. *IEEE Transactions on Knowledge and Data Engineering*, 17: 786–795, jun 2005. doi: http://dx.doi.org/10.1109/TKDE.2005.95.

Shan-Hung Wu, Keng-Pei Lin, Chung-Min Chen, and Ming-Syan Chen. Asymmetric support vector machines: low false-positive learning under the user tolerance. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '08*, pages 749–757, New York, New York, USA, aug 2008. ACM Press.

Xiaoyun Wu and Rohini K Srihari. New nu-Support Vector Machines and Their Sequential Minimal Optimization. In Tom Fawcett and Nina Mishra, editors, *Proceedings of the Twentieth International Conference (ICML 2003)*, pages 824–831. AAAI Press, 2003.

Bianca Zadrozny and Charles Elkan. Learning and making decisions when costs and probabilities are both unknown. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, pages 204–213, San Francisco, California, 2001. ACM. ISBN 1-58113-391-X. doi: 10.1145/502512.502540.

Bianca Zadrozny, John Langford, and Naoki Abe. Cost-sensitive learning by cost-proportionate example weighting. *Proceedings of the Third IEEE International Conference on Data Mining*, pages 435–442, 2003. doi: 10.1109/ICDM.2003.1250950.

Zhi-Qiang Zeng and Ji Gao. Improving SVM classification with imbalance data set. In *International Conference on Neural Information Processing*, pages 389–398. Springer, 2009.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32(1):56–85, 2004.