Sparse Reject Option Classifier using Successive Linear Programming

by

kulin Shah, Naresh Manwani

in

Thirty-Third AAAI Conference on Artificial Intelligence
(*AAAI-2019*)

Honolulu, Hawaii, USA

Report No: IIIT/TR/2019/-1

Centre for Cognitive Science
International Institute of Information Technology
Hyderabad - 500 032, INDIA
January 2019

Sparse Reject Option Classifier using Successive Linear Programming

Kulin Shah, Naresh Manwani
Machine Learning Lab, KCIS
IIIT, Hyderabad-500032

Abstract

In this paper, we propose an approach for learning sparse reject option classifiers using double ramp loss L_{dr}. We use DC programming to find the risk minimizer. The algorithm solves a sequence of linear programs to learn the reject option classifier. We show that the loss L_{dr} is Fisher consistent. We also show that the excess risk of loss L_d is upper bounded by excess risk of L_{dr}. We derive the generalization error bounds for the proposed approach. We show the effectiveness of the proposed approach by experimenting it on several real world datasets. The proposed approach not only performs comparable to the state of the art, it also successfully learns sparse classifiers.

1 Introduction

Standard classification tasks focus on building a classifier which predicts well on future examples. The overall goal is to minimize the number of mis-classifications. However, when the cost of mis-classification is very high, a generic classifier may still suffer from very high risk. In such cases it makes more sense not to classify high risk examples. This choice given to the classifier is called reject option. Hence, the classifiers which can also reject examples are called reject option classifiers. The rejection also has its cost but it is very less compared to the cost of mis-classification.

For example, making a poor decision based on the diagnostic reports can cost huge amount of money on further treatments or it can be cost of a life (da Rocha Neto et al. 2011). If the reports are ambiguous or some rare symptoms are seen which are unexplainable without further investigation, then the physician might choose not to risk misdiagnosing the patient. In this case, he might instead choose to perform further medical tests, or to refer the case to an appropriate specialist. Reject option classifier may also be found useful in financial services (Rosowsky and Smith 2013). Consider a banker looking at a loan application of a customer. He may choose not to decide on the basis of the information available, and ask for a credit bureau score or further recommendations from the stake holders. Reject option classifiers have been used in wide range of applications from healthcare (Hanczar and Dougherty 2008; da Rocha Neto et al. 2011) to text categorization (Fumera, Pillai, and Roli 2003) to crowd sourcing (Li et al. 2017) etc.

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Reject option classifier can be viewed as combination of a classifier and a rejection function. The rejection region impacts the proportion of examples that are likely to be rejected, as well as the proportion of predicted examples that are likely to be correctly classified. An optimal reject option classifier is the one which minimizes the rejection rate as well as the mis-classification rate on the predicted examples.

Let $\mathcal{X} \subseteq \mathbb{R}^n$ be the feature space and $\mathcal{Y}$ be the label space. For binary classification, we use $\mathcal{Y} = \{+1, -1\}$. Examples $(\mathbf{x}, y)$ are generated from unknown joint distribution $\mathcal{D}$ on the product space $\mathcal{X} \times \mathcal{Y}$. A typical *reject option classifier* is defined using a decision surface ($f(\mathbf{x}) = 0$) and bandwidth parameter ρ (determines rejection region) as follows:

$$h_\rho(f(\mathbf{x})) = 1.I_{\{f(\mathbf{x})>\rho\}} + 0.I_{\{|f(\mathbf{x})|\leq\rho\}} - 1.I_{\{f(\mathbf{x})<-\rho\}} \tag{1}$$

A reject option classifier can be viewed as two parallel surfaces and the area between them as rejection region. The goal is to determine both f and ρ simultaneously. The performance of a reject option classifier is measured using L_d loss function defined as:

$$L_d(yf(\mathbf{x}), \rho) = 1.I_{\{yf(\mathbf{x})<-\rho\}} + d.I_{\{|f(\mathbf{x})|\leq\rho\}} \tag{2}$$

where d is the cost of rejection. If $d = 0$, then $f(.)$ will always reject. If $d \geq 0.5$, then $f(\mathbf{x})$ will never reject, since the cost of random labeling is 0.5. Thus, d is chosen in the range $(0, 0.5)$. $h_\rho(f(\mathbf{x}))$ (described in equation. 1) has been shown to be infinite sample consistent with respect to the generalized Bayes classifier (Yuan and Wegkamp 2010). A reject option classifier is learnt by minimizing the risk which is the expectation of L_d with respect to the joint distribution $\mathcal{D}$. The risk under L_d is minimized by *generalized Bayes discriminant* $f_d^*(\mathbf{x})$ (Chow 1970), which is

$$f_d^*(\mathbf{x}) = 1.\mathbb{I}_{\{\eta(\mathbf{x})>1-d\}} + 0.\mathbb{I}_{\{d\leq\eta(\mathbf{x})\leq1-d\}} - 1.I_{\{\eta(\mathbf{x})<d\}} \tag{3}$$

where $\eta(\mathbf{x}) = P(y = 1|\mathbf{x})$. However, in general we do not know $\mathcal{D}$. But, we have the access to a finite set of examples drawn from $\mathcal{D}$ called training set. We find the reject option classifier by minimizing the empirical risk. Minimizing the empirical risk under L_d is computationally hard. To overcome this problem, convex surrogates of L_d have been

proposed. Generalized hinge based convex loss has been proposed for reject option classifier (Bartlett and Wegkamp 2008). The paper describes an algorithm for minimizing l_2 regularized risk under generalized hinge loss. Wegkamp et.al 2011 (Wegkamp and Yuan 2011) propose sparse reject option approach by minimizing l_1 regularized risk under generalized hinge loss. In both these approaches (Bartlett and Wegkamp 2008; Wegkamp and Yuan 2011), first a classifier is learnt based on risk minimization under generalized hinge loss and then a rejection threshold is learnt. Ideally, the classifier and the rejection threshold should be found simultaneously. This approach might not give the optimal parameters. Also, a very limited experimental results are provided to show the effectiveness of the proposed approaches (2011). A cost sensitive convex surrogate for L_d called double hinge loss has been proposed in (Grandvalet et al. 2008). The double hinge loss remains an upper bound to L_d provided $\rho \in \left(\frac{1-H(d)}{1-d}, \frac{H(d)-d}{d}\right)$, which is very strict condition. So far, the approaches proposed learn a threshold for rejection along with the classifier. However, in general, the rejection region may not be symmetrically located near the classification boundary. A generic convex approach has been proposed which simultaneously learns the classifier as well as the rejection function (Cortes, Salvo, and Mohri 2016). The main challenge with the convex surrogates is that they are not constant even in the reject region in contrast to L_d loss. Sousa and Cardoso (Sousa and Cardoso 2013) model reject option classification as ordinal regression problem. It is not clear whether treating rejection as a separate class leads to a good approximation simply because training data does not contain rejection as a class label. Moreover, classification consistency of this approach is not known in the reject option context. A non-convex formulation for learning reject option classifier using logistic function is proposed in Fumera and Roli (2002a). However, theoretical guarantees for the approach are not known. Also, a very limited set of experiments are provided in support of the approach. A bounded non-convex surrogate called *double ramp loss* L_{dr} is proposed in Manwani *et al.* (2015). A regularized risk minimization algorithm was proposed with l_2 regularization (Manwani et al. 2015). The approach proposed shown to have interesting geometric properties and robustness to the label noise. However, statistical properties of L_{dr} (Fisher consistency, generalization error etc.) are not studied so far. Also, l_2 regularization based approach does not learn sparse classifiers.

Our Contributions

In this paper, we propose a sparse reject option classifier learning algorithm using double ramp loss. By sparseness, we mean that the number of support vectors needed to express the classifier are small. Our contributions in this work are as follows.

- We propose a difference of convex (DC) programming (Thi Hoai An and Dinh Tao 1997) based algorithm to learn sparse reject option classifier. The final algorithm turns out to be solving successive linear programs.

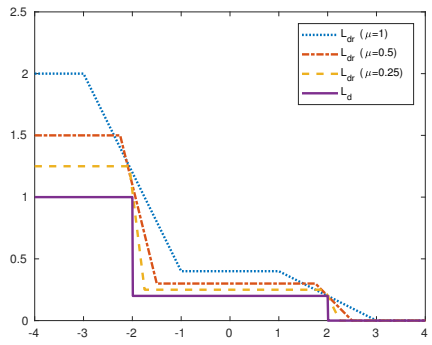

Figure 1: L_d vs. Double ramp loss L_{dr} (d=0.2, $\rho = 2$).

- We also establish statistical properties for double ramp loss function. We show that the double ramp loss function is Fisher consistent. Which means that generalized Bayes classifier minimizes the population risk under L_{dr}. We also show that the excess risk under loss L_{dr} upper bounds the excess risk under loss L_d.
- We derive the generalization error bounds for the proposed approach.
- We also show experimentally that the proposed approach performs comparable to the other state of the art approaches for reject option classifier. Our approach learns sparser classifiers compared to all the other approaches. We also show experimentally that the proposed approach is robust against label noise.

The rest of the paper is organized as follows. We discuss the proposed method and algorithm in section 2. In section 3, we provide the theoretical results for L_{dr}. The experiments are given in section 4. We conclude the paper with some remarks in section 5.

2 Proposed Approach

We propose a new algorithm for learning reject option classifier which minimizes the l_1-regularized risk under double ramp loss function L_{dr} (Manwani et al. 2015). L_{dr} is a non-convex surrogate of L_d as follows.

$$L_{dr}(t,\rho) = \frac{d}{\mu}\Big[[\mu - t + \rho]_+ - [-\mu^2 - t + \rho]_+\Big] + \frac{(1-d)}{\mu}\Big[[\mu - t - \rho]_+ - [-\mu^2 - t - \rho]_+\Big] \quad (4)$$

where μ is the slope of the loss in linear region, $[a]_+ = \max(0,a)$ and $t = yf(\mathbf{x})$. Note that L_{dr} depends on specific choice of μ. Also, for a valid reject region, we want $\rho \geq \frac{1}{2}\mu(1+\mu)$. Figure 1 shows the plot of L_{dr} for different values of μ.

Sparse Double Ramp SVM (SDR-SVM)

Let $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ be the training set where $(\mathbf{x}_i, y_i) \in \mathcal{X} \times \{+1, -1\},\ i = 1 \ldots N$. Let the reject option

classifier be of the form $f(\mathbf{x}) = h(\mathbf{x}) + b$. Let $\mathcal{K} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$ be a Mercer kernel (continuous, symmetric and positive semi-definite) to produce nonlinear classifiers. Let $\mathcal{H}_\mathcal{K}$ be the reproducing kernel Hilbert space (RKHS) induced by the Mercer kernel $\mathcal{K}$ with the norm $\|.\|_\mathcal{K}$ (Aronszajn 1950). To learn sparse reject option classifier, we use l_1 regularization term. Thus, we find the classifier as solving following optimization problem.

$$\min_{h \in \mathcal{H}^+_\mathcal{K}, b, \rho} \quad \lambda \|h\|_1 + \sum_{i=1}^{N} L_{dr}(y_i f(\mathbf{x}_i), \rho)$$

However, the optimal h lies in a finite dimensional subspace $\mathcal{H}^+_{\mathcal{K},S}$ of $\mathcal{H}_\mathcal{K}$ (Scholkopf and Smola 2001). $\mathcal{H}^+_{\mathcal{K},S} = \left\{ \sum_{i=1}^{N} y_i \alpha_i \mathcal{K}(\mathbf{x}_i, .) \mid [\alpha_1, \ldots, \alpha_N] \in \mathbb{R}^N_+ \right\}$. Given $h \in \mathcal{H}^+_{\mathcal{K},S}$, the l_1 regularization is defined as $\Omega(h) = \sum_{i=1}^{N} \alpha_i$ for $h(\mathbf{x}) = \sum_{i=1}^{N} y_i \alpha_i \mathcal{K}(\mathbf{x}_i, \mathbf{x})$ (Smola, Scholkopf, and Ratsch 1999; Bradley and Mangasarian 2000; Wu and Zhou 2005). Thus, the sparse double ramp SVM can be learnt by minimizing following l_1 regularized risk.

$$J(\Theta) = \lambda \sum_{i=1}^{N} \alpha_i + \frac{1}{N} \sum_{i=1}^{N} L_{dr}(y_i f(\mathbf{x}_i), \rho) \qquad (5)$$

where $f(\mathbf{x}_i) = \sum_{j=1}^{N} y_j \alpha_j \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) + b$. $\Theta = (\boldsymbol{\alpha}, b, \rho)$. We see that J is a non-convex function. However, J can be decomposed as a difference of two convex functions Q_1 and Q_2 as $J(\Theta) = Q_1(\Theta) - Q_2(\Theta)$, where

$$Q_1(\Theta) = \lambda \sum_{i=1}^{N} \alpha_i + \frac{1}{N\mu} \sum_{i=1}^{N} \Big[d\big[\mu - y_i f(\mathbf{x}_i) + \rho\big]_+ + (1-d)\big[\mu - y_i f(\mathbf{x}_i) - \rho\big]_+ \Big]$$

$$Q_2(\Theta) = \frac{1}{N\mu} \sum_{i=1}^{N} \Big[d\big[-\mu^2 - y_i f(\mathbf{x}_i) + \rho\big]_+ + (1-d)\big[-\mu^2 - y_i f(\mathbf{x}_i) - \rho\big]_+ \Big]$$

To minimize such a function which can be expressed as difference of two convex functions, we can use difference of convex (DC) programming. In this case, DC programming guarantees to find a local optima of the objective function (Thi Hoai An and Dinh Tao 1997). The simplified DC algorithm uses the convexity property of $Q_2(\Theta)$ and finds an upper bound on $J(\Theta)$ as $J(\Theta) \leq B(\Theta, \Theta^{(l)})$, where

$$B(\Theta, \Theta^{(l)}) := Q_1(\Theta) - Q_2(\Theta^{(l)}) - (\Theta - \Theta^{(l)})^T \nabla Q_2(\Theta^{(l)})$$

$\Theta^{(l)}$ is the parameter vector after $(l)^{th}$ iteration, $\nabla Q_2(\Theta^{(l)})$ is a sub-gradient of Q_2 at $\Theta^{(l)}$. $\Theta^{(l+1)}$ is found by minimizing $B(\Theta, \Theta^{(l)})$. Thus,

$$J(\Theta^{(l+1)}) \leq B(\Theta^{(l+1)}, \Theta^{(l)}) \leq B(\Theta^{(l)}, \Theta^{(l)}) = J(\Theta^{(l)})$$

Thus, the DC program reduces the value of $J(\Theta)$ in every iteration. Now, we will derive a DC algorithm for minimizing $J(\Theta)$. Given $\Theta^{(l)}$, we find

Algorithm 1 Sparse Double Ramp SVM (SDR-SVM)

Input: $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, $\epsilon > 0$, $d \in (0, 0.5)$, $\mu \in (0, 1]$, $\lambda > 0$
Output: $\boldsymbol{\alpha}^*, b^*, \rho^*$
Initialize: $l = 0$, $\boldsymbol{\alpha}^{(0)}, b^{(0)}, \rho^{(0)}$
while $(J(\Theta^{(l)}) - J(\Theta^{(l+1)}) > \epsilon)$ **do**
  for $i = 1$ to N **do**
    $\beta_i'^{(l)} = \mathbb{I}_{\{y_i f^{(l)}(\mathbf{x}_i) \leq \rho^{(l)} - \mu^2\}}$
    $\beta_i''^{(l)} = \mathbb{I}_{\{y_i f^{(l)}(\mathbf{x}_i) \leq -\rho^{(l)} - \mu^2\}}$
  end for
  $\boldsymbol{\alpha}^{(l+1)}, b^{(l+1)}, \rho^{(l+1)} = \arg\min_\Theta B(\Theta, \Theta^{(l)})$
end while

$\Theta^{(l+1)} \in \arg\min_\Theta B(\Theta, \Theta^{(l)}) = \arg\min_\Theta Q_1(\Theta) - \Theta^T \nabla Q_2(\Theta^{(l)})$. We use $\nabla Q_2(\Theta^{(l)})$ as:

$$\nabla Q_2(\Theta^{(l)}) = -\sum_{i=1}^{N} \begin{pmatrix} \frac{d\beta_i'^{(l)} + (1-d)\beta_i''^{(l)}}{\mu N} y_1 y_i \mathcal{K}(\mathbf{x}_1, \mathbf{x}_i) \\ \vdots \\ \frac{d\beta_i'^{(l)} + (1-d)\beta_i''^{(l)}}{\mu N} y_N y_i \mathcal{K}(\mathbf{x}_N, \mathbf{x}_i) \\ \frac{d\beta_i'^{(l)} + (1-d)\beta_i''^{(l)}}{\mu N} y_i \\ -\frac{d\beta_i'^{(l)} - (1-d)\beta_i''^{(l)}}{\mu N} \end{pmatrix}$$

where

$$\beta_i'^{(l)} = \mathbb{I}_{\{y_i f^{(l)}(\mathbf{x}_i) \leq \rho^{(l)} - \mu^2\}};\ i = 1 \ldots N$$
$$\beta_i''^{(l)} = \mathbb{I}_{\{y_i f^{(l)}(\mathbf{x}_i) \leq -\rho^{(l)} - \mu^2\}};\ i = 1 \ldots N$$

Note that $f^{(l)}(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i^{(l)} y_i \mathcal{K}(\mathbf{x}_i, \mathbf{x}) + b^{(l)}$. The new parameters $\Theta^{(l+1)}$ are found by minimizing $B(\Theta, \Theta^{(l)})$ subject to $\rho \geq \frac{1}{2}\mu(1+\mu)$. Which becomes

$$\min_{\boldsymbol{\alpha}, b, \rho, \boldsymbol{\xi}', \boldsymbol{\xi}''} \lambda \sum_{i=1}^{N} \alpha_i + \frac{1}{N\mu} \sum_{i=1}^{N} \big(d\xi_i' + (1-d)\xi_i'' \big) + \frac{d}{N\mu} \sum_{i=1}^{N} \beta_i'^{(l)} \big[y_i \big(\sum_{j=1}^{N} \alpha_j y_j \mathcal{K}(\mathbf{x}_j, \mathbf{x}_i) + b \big) - \rho \big] + \frac{1-d}{N\mu} \sum_{i=1}^{N} \beta_i''^{(l)} \big[y_i \big(\sum_{j=1}^{N} \alpha_j y_j \mathcal{K}(\mathbf{x}_j, \mathbf{x}_i) + b \big) + \rho \big]$$

$$s.t. \begin{cases} y_i \big(\sum_{j=1}^{N} \alpha_j y_j \mathcal{K}(\mathbf{x}_j, \mathbf{x}_i) + b \big) \geq \rho + \mu - \xi_i' \ \forall i \\ y_i \big(\sum_{j=1}^{N} \alpha_j y_j \mathcal{K}(\mathbf{x}_j, \mathbf{x}_i) + b \big) \geq -\rho + \mu - \xi_i'' \ \forall i \\ \alpha_i, \xi_i', \xi_i'' \geq 0 \ \ \forall i \quad \rho \geq \frac{1}{2}\mu(1+\mu) \end{cases}$$

Thus, $B(\Theta, \Theta^{(l)})$ can be minimized by solving a linear program. Thus, the algorithm solves a sequence of linear programs to learn a sparse reject option classifier. The complete approach is described in Algorithm 1. Convergence guarantee of this algorithm follows from the convergence of DC algorithm given in (Thi Hoai An and Dinh Tao 1997). The final learnt classifier is represented as $f(\mathbf{x}) = h(\mathbf{x}) + b$ and ρ.

3 Analysis

In this paper, we are proposing an algorithm based on L_{dr}. We first need to ensure that minimizer of the population risk under L_{dr} is minimized by the generalized Bayes classifier f_d^* (defined in eq.(3)). This property is called Fisher consistency or classification calibrated-ness.

Theorem 1. Fisher Consistency of L_{dr} *The generalized Bayes discriminant function* $f_d^*(\mathbf{x})$ *(described in eq. (3)) minimizes the risk*

$$\mathcal{R}_{dr}(f,\rho) = \mathbb{E}\big[L_{dr}(yf(\mathbf{x}),\rho)\big]$$

over all measurable functions f.

The proof of this theorem is skipped due to the space constraints and is provided in the supplementary file submitted with the main file. To approximate the optimal classifier, Fisher consistency is the minimal requirement for the loss function.

Excess Risk Bound

We will now derive the bound on the excess risk $(\mathcal{R}_d(f,\rho) - \mathcal{R}_d(f_d^*,\rho_d^*))$ in terms of the excess risk under L_{dr} where $\mathcal{R}_d(f,\rho) = \mathbb{E}[L_d(yf(\mathbf{x}),\rho)]$. We know that $L_d(f(\mathbf{x}),\rho) \leq L_{dr}(f(\mathbf{x}),\rho), \quad \forall \mathbf{x} \in \mathcal{X}, \quad \forall f$. This relation remains preserved when we take expectations both side, means $\mathcal{R}_d(f,\rho) \leq \mathcal{R}_{dr}(f,\rho)$. This relation is also true for excess risk. To show that, We first define the following terms. Let $\eta(\mathbf{x}) = P(y=1|\mathbf{x})$ and $z = f(\mathbf{x})$. We define following terms.

$$\begin{aligned}
\xi(\eta) &:= \eta\mathbb{I}_{\{\eta<d\}} + d\mathbb{I}_{\{d\leq\eta\leq 1-d\}} + (1-\eta)\mathbb{I}_{\{\eta>1-d\}} \\
H(\eta) &:= \inf_{z,\rho} \ \eta L_{dr}(z,\rho) + (1-\eta)L_{dr}(-z,\rho) \\
&= \eta(1+\mu)\mathbb{I}_{\{\eta<d\}} + d(1+\mu)\mathbb{I}_{\{d\leq\eta\leq 1-d\}} \\
&\quad + (1-\eta)(1+\mu)\mathbb{I}_{\{\eta>1-d\}}
\end{aligned}$$

We know that $\mathcal{R}_d^* = \mathbb{E}[\xi(\eta)]$ and $\mathcal{R}_{dr}^* = \mathbb{E}[H(\eta)]$. Furthermore, we define

$$\begin{aligned}
\xi_{-1}(\eta) &:= \eta - \xi(\eta) \\
\xi_r(\eta) &:= d - \xi(\eta) \\
\xi_1(\eta) &:= (1-\eta) - \xi(\eta) \\
H_{-1}(\eta) &:= \inf_{z<-\rho} \ \eta L_{dr}(z,\rho) + (1-\eta)L_{dr}(-z,\rho) \\
H_r(\eta) &:= \inf_{|z|\leq\rho} \ \eta L_{dr}(z,\rho) + (1-\eta)L_{dr}(-z,\rho) \\
H_1(\eta) &:= \inf_{z>\rho} \ \eta L_{dr}(z,\rho) + (1-\eta)L_{dr}(-z,\rho)
\end{aligned}$$

We observe the following relationship.

Proposition 2.

$$\begin{aligned}
\xi_{-1}(\eta) &\leq H_{-1}(\eta) - H(\eta) \\
\xi_r(\eta) &\leq H_r(\eta) - H(\eta) \\
\xi_1(\eta) &\leq H_1(\eta) - H(\eta)
\end{aligned}$$

The proof of the Proposition 2 is omitted due to the space constraints. Now we prove that the excess risk of L_d loss is bounded by excess risk of L_{dr} using above proposition.

Theorem 3. *For any measurable function* $f : \mathcal{X} \to \mathbb{R}$,

$$\mathcal{R}_d(f,\rho) - \mathcal{R}_d(f_d^*,\rho_d^*) \leq \mathcal{R}_{dr}(f,\rho) - \mathcal{R}_{dr}(f_d^*,\rho_d^*)$$

Proof. We know that

$$\mathcal{R}_d(f,\rho) = \mathbb{E}[\eta\mathbb{I}_{\{f<-\rho\}} + d\mathbb{I}_{\{-\rho\leq f\leq\rho\}} + (1-\eta)\mathbb{I}_{\{f>\rho\}}]$$

and $\mathcal{R}_{dr}(f,\rho) = \mathbb{E}[r_\eta(f)]$ where $r_\eta(f(\mathbf{x})) = \mathbb{E}_{y|\mathbf{x}}[L_{dr}(yf(\mathbf{x}),\rho)] = \eta L_{dr}(f(\mathbf{x}),\rho) + (1-\eta)L_{dr}(-f(\mathbf{x}),\rho)$. Therefore,

$$\begin{aligned}
&\mathcal{R}_d(f,\rho) - \mathcal{R}_d(f_d^*,\rho_d^*) \\
&= \mathbb{E}\big[\eta\mathbb{I}_{\{f<-\rho\}} + d\mathbb{I}_{\{|f|\leq\rho\}} + (1-\eta)\mathbb{I}_{\{f>\rho\}}\big] - \mathbb{E}\big[\xi(\eta)\big] \\
&= \mathbb{E}\big[\xi_{-1}(\eta)\mathbb{I}_{\{f<-\rho\}} + \xi_r(\eta)\mathbb{I}_{\{-\rho\leq f\leq\rho\}} + \xi_1(\eta)\mathbb{I}_{\{f>\rho\}}\big]
\end{aligned}$$

Using Proposition 2, we will get

$$\begin{aligned}
\mathcal{R}_d(f,\rho) - \mathcal{R}_d(f_d^*,\rho_d^*) \ &\leq \ \mathbb{E}\big[(H_{-1}(\eta) - H(\eta))\mathbb{I}_{\{f<-\rho\}} \\
&+ (H_r(\eta) - H(\eta))\mathbb{I}_{\{-\rho\leq f\leq\rho\}} + (H_1(\eta) - H(\eta))\mathbb{I}_{\{f>\rho\}}\big] \\
&\leq \ \mathbb{E}\big[H_{-1}(\eta)\mathbb{I}_{\{f<-\rho\}} + H_r(\eta)\mathbb{I}_{\{-\rho\leq f\leq\rho\}} \\
&\quad + H_1(\eta)\mathbb{I}_{\{f>\rho\}} - H(\eta)\big] \\
&\leq \ \mathbb{E}[r_\eta(f) - H(\eta)] \leq \ \mathcal{R}_{dr}(f,\rho) - \mathcal{R}_{dr}(f_d^*,\rho_d^*)
\end{aligned}$$

□

Hence, excess risk under L_d is upper bounded by excess risk under L_{dr}. From Theorem 3, we need to bound $\mathcal{R}_{dr}(f,\rho) - \mathcal{R}_{dr}(f_d^*,\rho_d^*)$ in order to bound $\mathcal{R}_d(f,\rho) - \mathcal{R}_d(f_d^*,\rho_d^*)$. We thus need an error decomposition for $\mathcal{R}_{dr}(f,\rho) - \mathcal{R}_{dr}(f_d^*,\rho_d^*)$.

Error Decomposition of $\mathcal{R}_{dr}(f,\rho) - \mathcal{R}_{dr}(f_d^*,\rho_d^*)$

The decomposition for RKHS based regularization schemes is well established (Cucker and Zhou 2007). To understand the details, consider the l_2 regularized empirical risk minimization with L_{dr}. For $S = \{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_N,y_N)\}$ and $\lambda_2 > 0$, let $f_{\lambda_2,S}^* = h_{\lambda_2,S}^* + b_{\lambda_2,S}^*$ where

$$(h_{\lambda_2,S}^*, b_{\lambda_2,S}^*, \rho_{\lambda_2,S}^*) = \arg\min_{h\in\mathcal{H}_\mathcal{K},b,\rho} \ \frac{\lambda_2}{2}\|h\|_\mathcal{K}^2 + \hat{\mathcal{R}}_{dr}(f,\rho) \tag{6}$$

Note that $\hat{\mathcal{R}}_{dr}$ denotes the empirical risk under double ramp loss. In this case, we observe the following decomposition.

$$\begin{aligned}
&\mathcal{R}_{dr}(f_{\lambda_2,S}^*,\rho_{\lambda_2,S}^*) - \mathcal{R}_{dr}(f_d^*,\rho_d^*) \leq \mathcal{A}(\lambda_2) + \mathcal{R}_{dr}(f_{\lambda_2,S}^*,\rho_{\lambda_2,S}^*) \\
&- \hat{\mathcal{R}}_{dr}(f_{\lambda_2,S}^*,\rho_{\lambda_2,S}^*) + \hat{\mathcal{R}}_{dr}(f_{\lambda_2}^*,\rho_{\lambda_2}^*) - \mathcal{R}_{dr}(f_{\lambda_2}^*,\rho_{\lambda_2}^*)
\end{aligned} \tag{7}$$

where $\hat{\mathcal{R}}_{dr}(f,\rho)$ is the empirical risk of (f,ρ) under double ramp loss. $f_{\lambda_2}^* = h_{\lambda_2}^* + b_{\lambda_2}^*$ and $\rho_{\lambda_2}^*$ are defined as follows.

$$(h_{\lambda_2}^*, b_{\lambda_2}^*, \rho_{\lambda_2}^*) = \arg\min_{h\in\mathcal{H}_\mathcal{K},b,\rho} \ \frac{\lambda_2}{2}\|h\|_\mathcal{K}^2 + \mathcal{R}_{dr}(f,\rho) \tag{8}$$

$\mathcal{A}(\lambda_2)$ measures the approximation power in RKHS $\mathcal{K}$ and is defined as follow.

$$\mathcal{A}(\lambda_2) = \inf_{h\in\mathcal{H}_\mathcal{K},b,\rho} \ \frac{\lambda_2}{2}\|h\|_\mathcal{K}^2 + \mathcal{R}_{dr}(h+b,\rho) - \mathcal{R}_{dr}(f_d^*,\rho_d^*) \quad \forall\lambda_2 > 0 \tag{9}$$

The error decomposition in eq.(7) is easy to derive once we know that both $h^*_{\lambda_2}$ and $h^*_{\lambda_2,S}$ lie in the same function space. However, this doe not hold true in case of SDR-SVM proposed in this paper. It happens because the error analysis becomes difficult due to the data dependent nature of $\mathcal{H}^+_{\mathcal{K}}$. We use the techniques discussed in (Wu and Zhou 2005; Huang, Shi, and Suykens 2014). We establish the error decomposition of SDR-SVM using the error decomposition (7) with the help of $f^*_{\lambda_2,S}$. We first characterize some properties of $f^*_{\lambda_2,S}, \rho^*_{\lambda_2,S}$. Note that from here onwards, we assume $\mu = 1$ (slope parameter in the loss function L_{dr}).

Proposition 4. *For any $\lambda_2 > 0$, $f^*_{\lambda_2,S} = (h^*_{\lambda_2,S}, b^*_{\lambda_2,S}, \rho^*_{\lambda_2,S})$ is given by eq.(6). Then,*

$$\Omega(h^*_{\lambda_2,S}) \leq \frac{1}{\lambda_2}\hat{\mathcal{R}}_{dr}(f^*_{\lambda_2,S}, \rho^*_{\lambda_2,S}) + \|h^*_{\lambda_2,S}\|^2_{\mathcal{K}}$$

The proof of this proposition is skipped here and is provided in the supplementary file.

Error Decomposition for SDR-SVM

We now find the error decomposition for SDR-SVM. We define the sample error as below,

$$\begin{aligned}\mathcal{S}(N, \lambda_1, \lambda_2) = &\big(\mathcal{R}_{dr}(f^*_{\lambda_1,S}, \rho^*_{\lambda_1,S}) - \hat{\mathcal{R}}_{dr}(f^*_{\lambda_1,S}, \rho^*_{\lambda_1,S})\big) \\ &+ (1+\psi)\big(\hat{\mathcal{R}}_{dr}(f^*_{\lambda_2}, \rho^*_{\lambda_2}) - \mathcal{R}_{dr}(f^*_{\lambda_2}, \rho^*_{\lambda_2})\big)\end{aligned}$$

where $(f^*_{\lambda_1,S}, \rho^*_{\lambda_1,S})$ is a global minimizer of optimization problem in eq.(5) and $(f^*_{\lambda_2}, \rho^*_{\lambda_2})$ is a global minimizer of problem (8). Also, $\psi = \frac{\lambda_1}{\lambda_2}$. Following theorem gives the error decomposition for SDR-SVM.

Theorem 5. *For $0 < \lambda_1 \leq \lambda_2 \leq 1$, let $\psi = \frac{\lambda_1}{\lambda_2}$. Then,*

$$\begin{aligned}&\mathcal{R}_{dr}(f^*_{\lambda_1,S}, \rho^*_{\lambda_1,S}) - \mathcal{R}_{dr}(f^*_d, \rho^*_d) + \lambda_1\Omega(h^*_{\lambda_1,S}) \\ &\quad\leq \psi\mathcal{R}_{dr}(f^*_d, \rho^*_d) + \mathcal{S}(N, \lambda_1, \lambda_2) + 2\mathcal{A}(\lambda_2)\end{aligned}$$

where $\mathcal{A}(\lambda_2)$ is the approximation error defined by eq.(9).

Proof of above theorem is provided in the supplementary file. Using Theorem 5, the generalization error of SDR-SVM is estimated by bounding $\mathcal{S}(N, \lambda_1, \lambda_2)$ and $\mathcal{A}(\lambda_2)$.

Generalization Error of SDR-SVM

We expect that the sample error $\mathcal{S}(N, \lambda_1, \lambda_2)$ tends to zero with certain rate as N tends to infinity. This can be understood by the convergence of the sample mean to its expected value. Also, we will have following assumption on $\mathcal{A}(\lambda_2)$.

Assumption 1. *For any $0 < \beta \leq 1$ and $c_\beta > 0$, the approximation error satisfies*

$$\mathcal{A}(\lambda_2) \leq c_\beta \lambda^\beta \quad \forall \lambda_2 > 0 \tag{10}$$

This is a standard assumption in the literature of learning theory (Cucker and Zhou 2007).

Theorem 6. *Suppose that Assumption 1 holds for any $0 < \beta \leq 1$. Take $\lambda_1 = N^{-\frac{\beta}{4\beta+2}}$ and $(f^*_{\lambda_1,S}, \rho^*_{\lambda_1,S})$ is the optimal solution of SDR-SVM. Then for any $0 \leq \delta \leq 1$, there holds*

$$\mathcal{R}_d(f^*_{\lambda_1,S}, \rho^*_{\lambda_1,S}) - \mathcal{R}_d(f^*_d, \rho^*_d) \leq \tilde{c}\left(log\frac{4}{\delta}\right)^{1/2} N^{-\frac{\beta}{4\beta+2}} \tag{11}$$

with probability at least $1-\delta$ where $\tilde{c} = 2c_\beta + 16d\tau^2 + 17$ and $\tau = \sup_{\mathbf{x},\mathbf{y}\in\mathcal{X}}\sqrt{|\mathcal{K}(\mathbf{x},\mathbf{y})|}$.

Proof of this theorem is provided in the supplementary file. It uses the concentration bounds results discussed in (Bartlett and Mendelson 2003).

Bounds with $\mu \neq 1$

Risk bound for $\mu \neq 1$ can be extended easily. The final expression for the risk bound ($\mu \neq 1$) is same as given in Theorem 6 with a different coefficient. The new coefficient $\tilde{c} = 2c_\beta + 13 + 4\max(1, \mu) + \frac{8d(1+\mu)\tau^2}{\mu}$.

4 Experiments

In this section, we show the effectiveness of approach on several datasets. We report experimental results on five datasets ("Ionosphere", "Parkinsons", "Heart", "ILPD" and "Pima Indian Diabetes") available on UCI machine learning repository (Lichman 2013).

Experimental Setup

In the proposed approach, to solve linear programming problem in each iteration, we have used CVXOPT package in python language (Dahl and Vandenberghe 2008). In our experiments, we apply a Gaussian kernel $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2)$ for nonlinear problems. In all the experiments, we set μ = 1. Regularization parameter λ and kernel parameter γ are chosen using 10-fold cross validation.

We compare the performance of the proposed approach (SDR-SVM) with 3 other approaches as follows. The first approach is standard SVM based reject option classifier. In that approach, we first learn a learning decision boundary using SVM and then set the width of rejection region by cross-validation such that empirical risk under $L_{d,\rho}$ is minimized. We use this approach as a proxy for the approach proposed in Bartlett and Wegkamp (2008). Again, parameters of SVM (C and γ) are learnt using 10-fold cross-validation. The second approach is the SVM with embedded reject option (ER-SVM) (Fumera and Roli 2002a). We used the code for this approach available online (Fumera and Roli 2002b). We also compare our approach with Double hinge SVM (DH-SVM) based reject option classifier (Grandvalet et al. 2008).

Simulation Results

We report the experimental results for different values of $d \in [0.05, 0.5]$ with the step size of 0.05. For every value of d, we find the cross-validation risk (under $L_{d,\rho}$), % rejection rate (RR), % accuracy on the un-rejected examples (Acc). We also report the average number of support vectors (the corresponding $\alpha_i \geq 10^{-6}$). The results provided here are based on 10 repetitions of 10-fold cross-validation (CV).

Now we discuss the experimental results. Figure 2 shows the comparison plots for different datasets. We observe the following.

1. **Average Cross Validation Risk $\mathcal{R}_d$:** We see that SDR-SVM performs better than ER-SVM with huge gaps in terms of the average cross validation risk ($\hat{\mathcal{R}}_d$) for all

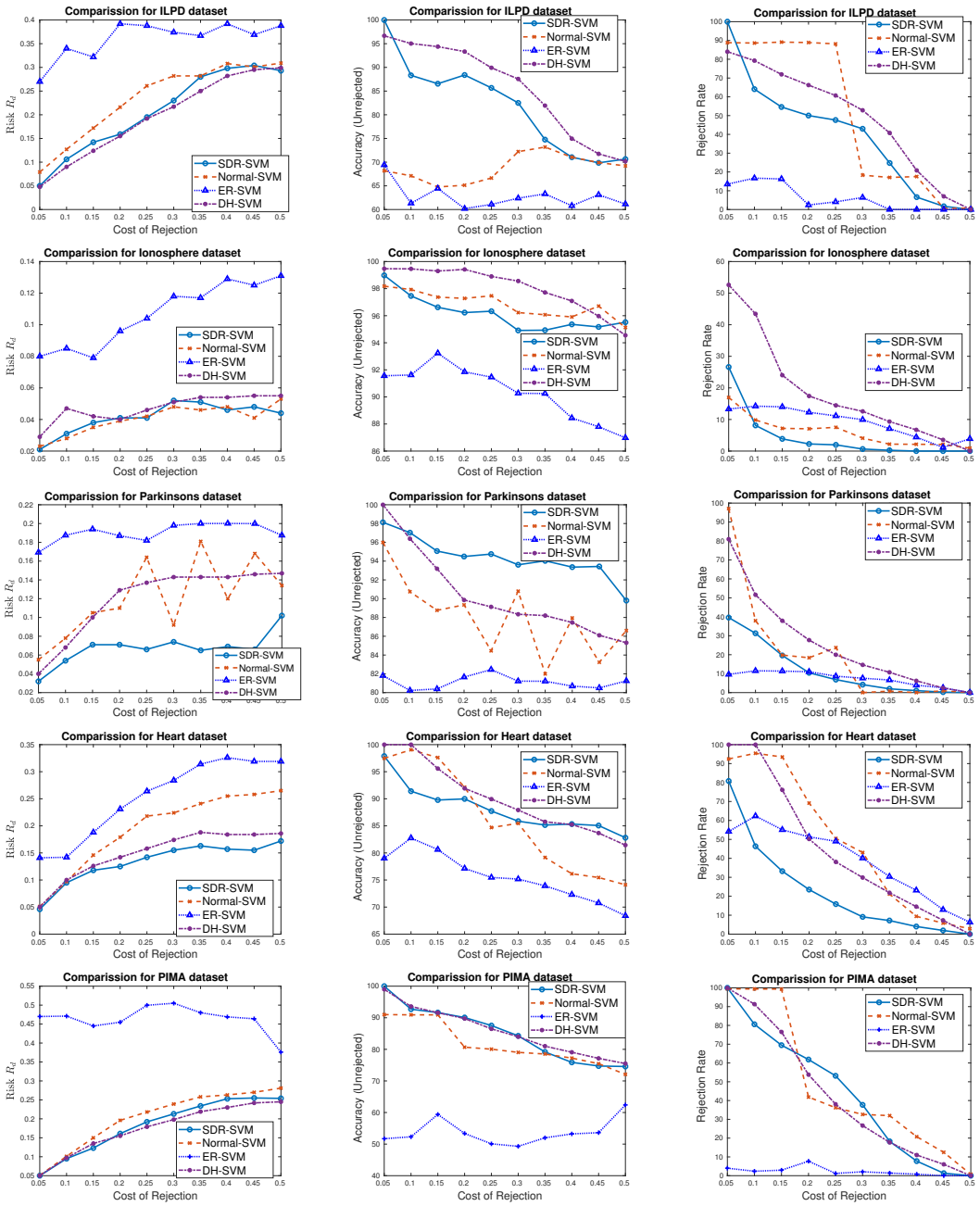

Figure 2: Comparison Plots for Different Datasets. Column 1 shows the risk R_d, column 2 shows accuracy on un-rejected examples, column 3 shows the rejection rate.

datasets and for all values of d. For Parkinsons and Heart datasets, SDR-SVM has smaller $\hat{\mathcal{R}}_d$ risk (for all values of d) compared to DH-SVM. For ILPD, Ionosphere and PIMA datasets, $\hat{\mathcal{R}}_d$ risk of SDR-SVM is comparable to DH-SVM. SDR-SVM performs better than Normal-SVM based approach on Parkinsons, Heart, ILPD and PIMA datasets. For Ionosphere dataset, SDR-SVM performs comparable to Normal-SVM based approach.

2. **Rejection Rate:** We observe that for Inosphere, Heart and Parkinsons datasets, rejection rate of SDR-SVM is much smaller compared to other approaches except for smaller values of d (0.05 and 0.1). For PIMA and ILPD datasets, the rejection rates of SDR-SVM are comparable to DH-SVM. The rejection rates for these two datasets are comparatively higher for all values of d. Possible reason for that could be high overlap between the two class regions.

3. **Performance on Unrejected Examples:** We see that

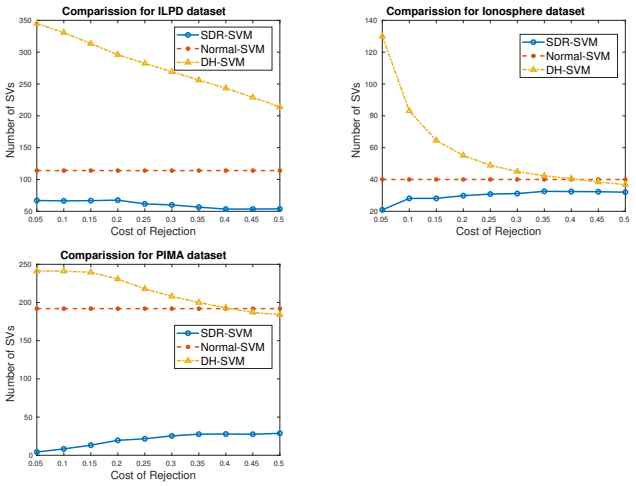

Figure 3: Sparseness Comparison of SDR-SVM with DH-SVM and Normal-SVM

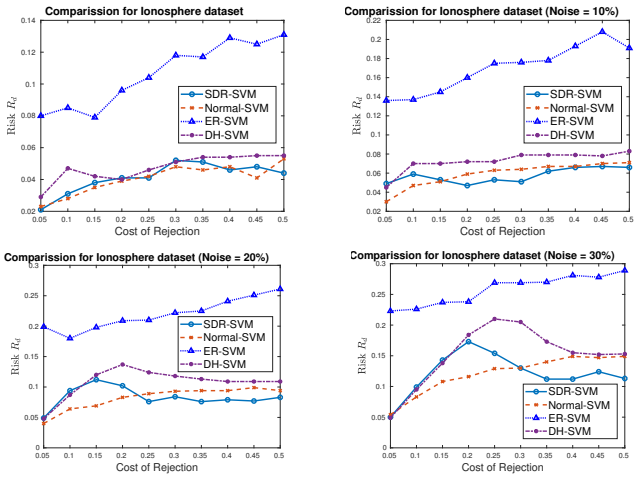

Figure 4: Comparison Results in presence of uniform Label Noise

SDR-SVM also gives good classification accuracy on unrejected examples. It always gives better accuracy compared ER-SVM. As compared to normal SVM based approach, SDR-SVM does always better on ILDP and Parkinsons datasets. For rest of the datasets, SDR-SVM gives comparable accuracy to normal SVM based method on unrejected examples. Compared to double hinge SVM, SDR-SVM does comparable to DH-SVM.

Thus, overall SDR-SVM learns reject option classifiers which attain smaller $\hat{\mathcal{R}}_d$ risk. It achieves this goal by simultaneously minimizing the rejection rate and misclassification rate on unrejected examples.

Sparseness Results

We now show that SDR-SVM learns sparse reject option classifiers. As discussed, by sparseness we mean that the resulting classifier can be represented as a linear combination of a very small fraction of training points. Sparseness results for SDR-SVM are shown in Figure 3.

We see that for ILPD, Ionosphere and PIMA datasets, SDR-SVM outputs classifiers which are much sparser compared to DH-SVM and Normal-SVM based approaches. ER-SVM does not have obvious representation for the classifier as a linear combination of training examples.

Experiments with Noisy Data

$L_{dr,\rho}$ is generalization of ramp loss function for the reject option classification. For normal binary classification problem, ramp loss function is shown robust against label noise (Ghosh, Manwani, and Sastry 2015). Motivated by the above fact, we did experiments to test the robustness of $L_{dr,\rho}$ against uniform label noise (with noise rates of $10\%, 20\%, 30\%$). Figure 4. We observe the following.

1. We observe that with 10% noise rate, increment in the risk for SDR-SVM is not significant. As we increase the noise rate, model in reject option classification confuses more for classifying the examples, therefore model tries to put more examples in rejection region for smaller values of d. Which leads to increase in width of rejection region. Thus, for smaller values of d, risk is dominated by rejection cost for proposed approach. But as we increase d, cost of rejection also increases and model in label noise will force examples to classify to one of the label. With increasing noise rate, SDR-SVM remains robust for higher values of d.
2. Compared to ER-SVM, SDR-SVM does significantly better for all values of d and for all noise rates.
3. For large values of d, SDR-SVM performs better than DH-SVM and noraml SVM in presence of label noise.

5 Conclusions

In this paper, we proposed sparse approach for learning reject option classifier using double ramp loss. We propose a DC programming based approach for minimizing the regularized risk. The approach solves successive linear programs to learn the classifier. Our approach also learns nonlinear classifier by using appropriate kernel function. Further, we have shown the Fisher consistency of double ramp loss $L_{dr,\rho}$. We upper bound the excess risk of L_d in terms of excess risk of L_{dr}. We then derive generalization bounds for SDR-SVM. We showed experimentally that the proposed approach does better compared to the other approaches for reject option classification and learns sparse classifiers. We also experimental evidences to show robustness of SDR-SVM against the label noise.

References

Aronszajn, N. 1950. Theory of reproducing kernels. *Transactions of the American mathematical society* 68(3):337–404.

Bartlett, P. L., and Mendelson, S. 2003. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* 3:463–482.

Bartlett, P. L., and Wegkamp, M. H. 2008. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research.* 9:1823–1840.

Bradley, P., and Mangasarian, O. 2000. Massive data discrimination via linear support vector machines. *Optimization methods and software* 13(1):1–10.

Chow, C. 1970. On optimum recognition error and reject tradeoff. *IEEE Trans. Inf. Theor.* 16(1):41–46.

Cortes, C.; Salvo, G. D.; and Mohri, M. 2016. Learning with rejection. In *Proceedings of 27th International Conference on Algorithmic Learning Theory (ALT)*, 67–82.

Cucker, F., and Zhou, D. X. 2007. *Learning Theory: An Approximation Theory Viewpoint (Cambridge Monographs on Applied & Computational Mathematics)*. Cambridge University Press.

da Rocha Neto, A. R.; Sousa, R.; de A. Barreto, G.; and Cardoso, J. S. 2011. Diagnostic of pathology on the vertebral column with embedded reject option. In *Pattern Recognition and Image Analysis*, 588–595.

Dahl, J., and Vandenberghe, L. 2008. CVXOPT: A python package for convex optimization.

Fumera, G., and Roli, F. 2002a. Support vector machines with embedded reject option. In *Pattern Recognition with Support Vector Machines: First International Workshop, SVM 2002 Niagara Falls, Canada, August 10, 2002 Proceedings*, 68–82.

Fumera, G., and Roli, F. 2002b. *Support Vector Machines with Embedded Reject Option*. `https://github.com/rjgsousa/RejectOption/tree/master/Fumera/fumera_code`.

Fumera, G.; Pillai, I.; and Roli, F. 2003. Classification with reject option in text categorisation systems. In *12th International Conference on Image Analysis and Processing, 2003.Proceedings.*, 582–587.

Ghosh, A.; Manwani, N.; and Sastry, P. 2015. Making risk minimization tolerant to label noise. *Neurocomput.* 160(C):93–107.

Grandvalet, Y.; Rakotomamonjy, A.; Keshet, J.; and Canu, S. 2008. Support vector machines with a reject option. In *Advances in Neural Information Processing Systems (NIPS)*, 537–544.

Hanczar, B., and Dougherty, E. R. 2008. Classification with reject option in gene expression data. *Bioinformatics* 24(17):1889–1895.

Huang, X.; Shi, L.; and Suykens, J. A. 2014. Ramp loss linear programming support vector machine. *Journal of Machine Learning Research* 15:2185–2211.

Li, Q.; Vempaty, A.; Varshney, L.; and Varshney, P. 2017. Multi-object classification via crowdsourcing with a reject option. *IEEE Transactions on Signal Processing* 65(4):1068–1081.

Lichman, M. 2013. UCI machine learning repository.

Manwani, N.; Desai, K.; Sasidharan, S.; and Sundararajan, R. 2015. Double ramp loss based reject option classifier. In *Advances in Knowledge Discovery and Data Mining - 19th Pacific-Asia Conference, PAKDD*, 151–163.

Rosowsky, Y. I., and Smith, R. E. 2013. Rejection based support vector machines for financial time series forecasting. In *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, 1161–1167.

Scholkopf, B., and Smola, A. J. 2001. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

Smola, A.; Scholkopf, B.; and Ratsch, G. 1999. Linear programs for automatic accuracy control in regression. In *Ninth International Conference on Artificial Neural Networks (ICANN)*, volume 2, 575–580.

Sousa, R., and Cardoso, J. S. 2013. The data replication method for the classification with reject option. *AI Commun.* 26(3):281–302.

Thi Hoai An, L., and Dinh Tao, P. 1997. Solving a class of linearly constrained indefinite quadratic problems by d.c. algorithms. *Journal of Global Optimization* 11(3):253–285.

Wegkamp, M., and Yuan, M. 2011. Support vector machines with a reject option. *Bernoulli* 17(4):1368–1385.

Wu, Q., and Zhou, D.-X. 2005. Svm soft margin classifiers: Linear programming versus quadratic programming. *Neural Comput.* 17(5):1160–1187.

Yuan, M., and Wegkamp, M. 2010. Classification methods with reject option based on convex risk minimization. *J. Mach. Learn. Res.* 11:111–130.