Classification with Rejection Based on Cost-sensitive Classification

Nontawat Charoenphakdee [1 2] **Zhenghang Cui** [1 2] **Yivan Zhang** [1 2] **Masashi Sugiyama** [2 1]

Abstract

The goal of classification with rejection is to avoid risky misclassification in error-critical applications such as medical diagnosis and product inspection. In this paper, based on the relationship between classification with rejection and cost-sensitive classification, we propose a novel method of classification with rejection by learning an ensemble of cost-sensitive classifiers, which satisfies all the following properties: (i) it can avoid estimating class-posterior probabilities, resulting in improved classification accuracy, (ii) it allows a flexible choice of losses including non-convex ones, (iii) it does not require complicated modifications when using different losses, (iv) it is applicable to both binary and multiclass cases, and (v) it is theoretically justifiable for any classification-calibrated loss. Experimental results demonstrate the usefulness of our proposed approach in clean-labeled, noisy-labeled, and positive-unlabeled classification.

1. Introduction

In ordinary classification, a classifier learned from training data is expected to accurately predict a label of every possible test input in the input space. However, when a particular test input is difficult to classify, forcing a classifier to always predict a label can lead to misclassification, causing serious troubles in risk-sensitive applications such as medical diagnosis, home robotics, and product inspection (Cortes et al., 2016a; Geifman & El-Yaniv, 2017; Ni et al., 2019). To cope with this problem, classification with rejection was proposed as a learning framework to allow a classifier to abstain from making a prediction (Chow, 1957; 1970; Bartlett & Wegkamp, 2008; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; Cortes et al., 2016a;b; Yuan & Wegkamp, 2010; Franc & Prusa, 2019), so that we can prevent misclassification in critical applications.

A well-known framework for classification with rejection that has been studied extensively is called the cost-based framework (Chow, 1970; Bartlett & Wegkamp, 2008; Yuan & Wegkamp, 2010; Cortes et al., 2016a;b; Franc & Prusa, 2019; Ni et al., 2019). In this setting, we set a pre-defined rejection cost to be less than the misclassification cost. As a result, a classifier trained in this framework prefers to reject than making a risky prediction, where there currently exist two main approaches as the following.

The first approach is called the *confidence-based approach*, where we train a classifier then use an output of the classifier as a confidence score (Bartlett & Wegkamp, 2008; Grandvalet et al., 2009; Herbei & Wegkamp, 2006; Yuan & Wegkamp, 2010; Ramaswamy et al., 2018; Ni et al., 2019). In this approach, we manually set a confidence threshold as a criterion to refrain from making a prediction, if the confidence score of a test input is lower than the threshold. Most confidence-based methods rely on a loss that can estimate class-posterior probabilities (Yuan & Wegkamp, 2010; Reid & Williamson, 2010; Ni et al., 2019), which can be difficult to estimate especially when using deep neural networks (Guo et al., 2017). Although there are some exceptions that can avoid estimating class-posterior probabilities, most of them are only applicable to binary classification (Bartlett & Wegkamp, 2008; Grandvalet et al., 2009; Manwani et al., 2015).

The second approach is called the *classifier-rejector* approach, where we simultaneously train a classifier and a rejector (Cortes et al., 2016a;b; Ni et al., 2019). It is known that this approach has theoretical justification in the binary case only for the exponential and hinge-based losses (Cortes et al., 2016a;b). This is because the proof technique highly relies on the function form of the loss (Cortes et al., 2016a;b). In the multiclass case, Ni et al. (2019) argued that this approach is not suitable both theoretically and experimentally since the multiclass extension of Cortes et al. (2016b) is not calibrated and the confidence-based softmax cross-entropy loss can outperform this approach in practice.

The goal of this paper is to develop an alternative approach to classification with rejection that achieves the following four design goals. First, it can avoid estimating class-posterior probabilities, since this often yields degradation

[1]The University of Tokyo, Tokyo, Japan [2]RIKEN AIP, Tokyo, Japan. Correspondence to: Nontawat Charoenphakdee <nontawat@ms.k.u-tokyo.ac.jp>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

of classification performance. Second, the choice of losses is flexible and does not require complicated modifications when using different losses, which allows a wider range of applications. Third, it is applicable to both binary and multiclass cases. Fourth, it can be theoretically justified. In this paper, we show that this goal can be achieved by bridging the theory of cost-sensitive classification (Elkan, 2001; Scott, 2012; Steinwart, 2007) and classification with rejection. The key observation that allows us to connect the two problems is based on the fact that one can mimic the Bayes optimal solution of classification rejection by only knowing $\arg\max_y p(y|x)$ and whether $\max_y p(y|x) > 1 - c$, where c is the rejection cost. Based on this observation, we propose the *cost-sensitive approach*, which calibration can be guaranteed for *any classification-calibrated loss* (Zhang, 2004; Bartlett et al., 2006). Classification-calibration is known to be a minimum requirement for a loss in ordinary classification (Bartlett et al., 2006). This suggests that the loss choices of our proposed approach are *as flexible as that of ordinary classification*.

To emphasize the importance of having a flexible loss choice, we explore the usage of our approach for classification from positive and unlabeled data (PU-classification) (du Plessis et al., 2014; 2015; Kiryo et al., 2017) and classification from noisy labels (Angluin & Laird, 1988; Ghosh et al., 2015). Our experimental results show that a family of symmetric losses, which are the losses that cannot estimate class-posterior probabilities (Charoenphakdee et al., 2019), can be advantageous in these settings. We also provide experimental results of clean-labeled classification with rejection to illustrate the effectiveness of the cost-sensitive approach.

2. Preliminaries

In this section, we introduce the problem setting of classification with rejection. Then, we review cost-sensitive binary classification, which will be essential for deriving the proposed cost-sensitive approach for classification with rejection.

2.1. Classification with Rejection

Our problem setting follows the standard cost-based framework classification with rejection (Chow, 1970; Cortes et al., 2016b; Ni et al., 2019). Let $\mathcal{X}$ be an input space and $\mathcal{Y} = \{1, \dots, K\}$ be an output space, where K denotes the number of classes. Note that we adopt a conventional notation $\mathcal{Y} = \{-1, +1\}$ when considering binary classification (Bartlett et al., 2006). In this problem, we are given the training input-output pairs $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ drawn i.i.d. from an unknown probability distribution with density $p(\boldsymbol{x}, y)$. A classification rule of learning with rejection is $f\colon \mathcal{X} \to \{1, \dots, K, ®\}$, where ® denotes rejection. Let $c \in (0, 0.5)$ be the rejection cost. Unlike ordinary classification, where the zero-one loss $\ell_{01}(f(\boldsymbol{x}), y) = \mathbb{1}_{[f(\boldsymbol{x}) \neq y]}$[1] is the performance measure, we are interested in an extension of ℓ_{01}, which is called the zero-one-c loss ℓ_{01c} defined as follows (Ni et al., 2019):

$$\ell_{01c}(f(\boldsymbol{x}), y) = \begin{cases} c & f(\boldsymbol{x}) = ®, \\ \ell_{01}(f(\boldsymbol{x}), y) & \text{otherwise.} \end{cases}$$

The goal is to find a classification rule f that minimizes the expected risk with respect to ℓ_{01c}, i.e.,

$$R^{\ell_{01c}}(f) = \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim p(\boldsymbol{x},y)} [\ell_{01c}(f(\boldsymbol{x}), y)]. \tag{1}$$

In classification with rejection, a classification rule f is allowed to refrain from making a prediction and will receive a fixed rejection loss c. In this paper, following most existing studies (Cortes et al., 2016a;b; Ramaswamy et al., 2018; Ni et al., 2019), we consider the case where $c < 0.5$. Intuitively, this case implies that it is strictly better to reject if a classifier has less than half a chance to be correct. Thus, the case where $c < 0.5$ is suitable if the goal is to avoid harmful misclassification. We refer the readers to Ramaswamy et al. (2018) for more discussion on the case where $c \geq 0.5$ and how it is fundamentally different from the case where $c < 0.5$. Next, let us define $\boldsymbol{\eta}(\boldsymbol{x}) = [\eta_1(\boldsymbol{x}), \dots, \eta_K(\boldsymbol{x})]^\top$, where $\eta_y(x) = p(y|\boldsymbol{x})$ denotes the class-posterior probability of a class y. The optimal solution for classification with rejection $f^* = \arg\min_f R^{\ell_{01c}}(f)$ known as Chow's rule (Chow, 1970) can be expressed as follows:

Definition 1 (Chow's rule (Chow, 1970))**.** The optimal solution of multiclass classification with rejection $f^* = \arg\min_f R^{\ell_{01c}}(f)$ can be expressed as

$$f^*(\boldsymbol{x}) = \begin{cases} ® & \max_y \eta_y(\boldsymbol{x}) \leq 1 - c, \\ \arg\max_y \eta_y(\boldsymbol{x}) & \text{otherwise.} \end{cases}$$

Chow's rule suggests that classification with rejection is solved if we have the knowledge of $\boldsymbol{\eta}(\boldsymbol{x})$. Therefore, one approach is to estimate $\boldsymbol{\eta}(\boldsymbol{x})$ from training examples. This method is in a family of the confidence-based approach, which has been extensively studied in both the binary (Yuan & Wegkamp, 2010) and multiclass cases (Ni et al., 2019). Figure 1 illustrates the confidence-based approach.

2.2. Cost-sensitive Binary Classification

Consider binary classification where $y \in \{-1, +1\}$. In ordinary classification, the false positive and false negative costs are treated equally. On the other hand, in cost-sensitive classification, the false positive and false negative costs are

[1] $\mathbb{1}_{[\cdot]}$ denotes an indicator function.

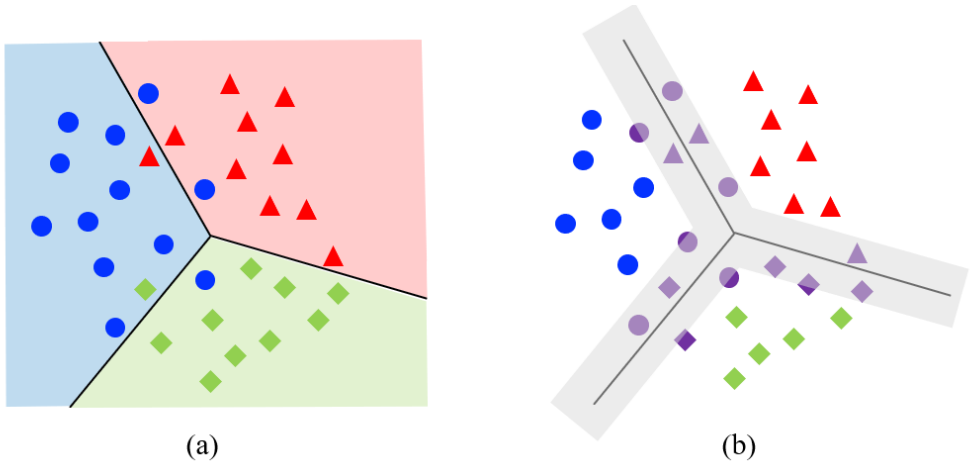

Figure 1. Illustration of the confidence-based approach. Figure (a) denotes a prediction function. The rejector in figure (b) has a rejection region spreads from the decision boundary of the prediction function. The width of the rejection region depends on the choice of the rejection threshold parameter. Data points in purple are rejected.

generally unequal (Elkan, 2001; Saerens et al., 2002; Scott, 2012).

Without loss of generality, we define $\alpha \in (0, 1)$ to be the false positive cost and $1 - \alpha$ to be the false negative cost. Then, the expected cost-sensitive risk can be expressed as

$$R_\alpha^{\ell_{01}}(f) = (1-\alpha)\pi \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p(x|y=+1)}[\ell_{01}(f(\boldsymbol{x}), +1)] + \alpha(1-\pi) \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p(x|y=-1)}[\ell_{01}(f(\boldsymbol{x}), -1)],$$

where $\pi = p(y = +1)$ denotes the class prior.

It is known that the Bayes optimal cost-sensitive binary classifier can be expressed as follows:

Definition 2 (Scott (2012))**.** The optimal cost-sensitive classifier $f_\alpha^* = \arg\min_f \mathbb{R}_\alpha(f)$ can be expressed as

$$f_\alpha^*(\boldsymbol{x}) = \begin{cases} +1 & p(y=+1|\boldsymbol{x}) > \alpha, \\ -1 & \text{otherwise.} \end{cases}$$

Note that when $\alpha = 0.5$, the Bayes optimal solution $f_{0.5}^*(\boldsymbol{x})$ coincides with that of ordinary binary classification. Moreover, when α is known, cost-sensitive binary classification is solved if we have access to the class-posterior probability $p(y = +1|\boldsymbol{x})$.

3. Cost-sensitive Approach

In this section, we propose a cost-sensitive approach for classification with rejection. We begin by describing our motivation and analyzing the behavior of the Bayes optimal solution of classification with rejection. Then, we show that this problem can be solved by simultaneously solving multiple cost-sensitive classification problems.

3.1. Motivation

As suggested by Chow's rule (Chow, 1970), classification with rejection can be solved by estimating the class-posterior probabilities. However, an important question arises as: *Is class-posterior probability estimation indispensable for solving classification with rejection?* This question is fundamentally motivated by Vapnik's principle (Vapnik, 1998), which suggests not to solve a more general problem as an intermediate step when solving a target problem if we are given a restricted amount of information.

In our context, the general problem is class-posterior probability estimation. In fact, knowing class-posterior probabilities can also solve many other problems (Qin, 1998; Bickel et al., 2007; Sugiyama et al., 2012; Dembczynski et al., 2013; Koyejo et al., 2014). However, many of such problems are also known to be solvable without estimating the class-posterior probabilities (Kanamori et al., 2012; Bao & Sugiyama, 2020). Note that class-posterior probability estimation can be unreliable when the model is misspecified (Begg & Lagakos, 1990; Heagerty & Kurland, 2001) or highly flexible (Guo et al., 2017; Hein et al., 2019).

To find a more direct solution for classification with rejection, we seek for a general approach that it may not be able to estimate class-posterior probabilities, but its optimal solution coincides with the optimal Chow's rule (Chow, 1970). Although the idea of directly solving classification with rejection without class-posterior estimation itself is not novel, most existing methods are only applicable to binary classification with rejection (Bartlett & Wegkamp, 2008; Grandvalet et al., 2009; Manwani et al., 2015; Cortes et al., 2016b;a), or focus on specific types of losses (Ramaswamy et al., 2018). For the multiclass case, Zhang et al. (2018)

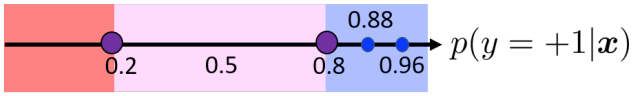

Figure 2. Illustration of Chow's rule in binary classification with rejection and the unnecessity of knowing the class-posterior probability to solve this problem. If the rejection cost $c = 0.2$, as long as we know $p(y = 1|x) > 0.8$, knowing the exact value of the class-posterior probability does not change our final decision to predict a positive label.

proposed to modify a loss by bending it to be steeper (for the hinge loss) and positively unbounded, but there exist hyperparameters to be tuned such as the rejection threshold and the bending slope. Also, Mozannar & Sontag (2020) recently proposed a classifier-rejector approach by augmenting a rejection class in the model's prediction, but their loss choice is limited to the modified cross-entropy loss. More discussion on related work is provided in Appendix A.

3.2. A Closer Look at Chow's Rule

Here, we analyze the behavior of Chow's rule (Chow, 1970). We discuss the minimal knowledge required for a classification rule to mimic Chow's rule, which illustrates that the class-posterior probabilities need not to be known. For simplicity, we begin by considering binary classification with rejection.

In binary classification with rejection, Chow's rule in Definition 1 can be expressed as

$$f^*(\boldsymbol{x}) = \begin{cases} 1 & p(y = +1|x) > 1 - c, \\ ® & c \leq p(y = +1|x) \leq 1 - c, \\ -1 & p(y = +1|x) < c. \end{cases} \tag{2}$$

To solve binary classification with rejection, there are only three conditions to verify, which are $p(y = +1|x) > 1 - c$, $p(y = +1|x) < c$, and $p(y = +1|x) \in [c, 1 - c]$. We can see that if we know $p(y = +1|x) > 1-c$, we do not need to know the exact value of $p(y = +1|x)$ to predict the label as positive. For example, if $c = 0.2$, knowing $p(y = +1|x) > 0.8$ is already sufficient to predict a label, i.e., knowing whether $p(y = +1|x) > 0.88$ or $p(y = +1|x) > 0.96$ does not change the decision of Chow's rule. Figure 2 illustrates this fact, which is the key intuition why it is possible to develop a method that can avoid estimating the class-posterior probabilities for solving this problem.

3.3. Binary Classification with Rejection Based on Cost-sensitive Classification

Here, we show that by solving two cost-sensitive binary classification problems, binary classification with rejection can be solved. The following proposition shows the relationship between the optimal solutions of cost-sensitive binary classification and that of binary classification with rejection.

Proposition 3. *In binary classification with rejection, Chow's rule can be expressed as*

$$f^*(\boldsymbol{x}) = \begin{cases} 1 & f^*_{1-c}(\boldsymbol{x}) = 1, \\ -1 & f^*_c(\boldsymbol{x}) = -1, \\ ® & \text{otherwise}. \end{cases} \tag{3}$$

Proof. We assert that if we can verify the following two conditions:

$$p(y = +1|x) > 1 - c, \tag{4}$$

$$p(y = +1|x) > c, \tag{5}$$

then binary classification with rejection is solved. Based on Chow's rule (2), if Ineq. (4) holds, Ineq. (5) must also hold since $c < 0.5$. Then we should predict a positive label. On the other hand, we should predict a negative label if Ineq. (5) does not hold. Next, if Ineq. (5) holds but Ineq. (4) does not hold, we should reject $\boldsymbol{x}$. As a result, based on Definition 2, knowing $f^*_{1-c}(\boldsymbol{x})$ and $f^*_c(\boldsymbol{x})$ is sufficient to verify Ineqs. (4) and Ineq. (5). This concludes the proof. □

Proposition 3 suggests that by solving two binary cost-sensitive classification with $\alpha = c$ and $\alpha = 1 - c$ to obtain $f^*_c(\boldsymbol{x})$ and $f^*_{1-c}(\boldsymbol{x})$, binary classification with rejection can be solved.

3.4. Multiclass Extension

Here, we show that our result in Section 3.3 can be naturally extended to the multiclass case. More specifically, we show that multiclass classification with rejection can be solved by learning an ensemble of K binary cost-sensitive classifiers.

Let us define the Bayes optimal solution for one-versus-rest cost-sensitive binary classifier $f^{*,y}_\alpha$, where y is the positive class and $y' \in \mathcal{Y}, y' \neq y$ are the negative classes:

$$f^{*,y}_\alpha(\boldsymbol{x}) = \begin{cases} +1 & \eta_y(\boldsymbol{x}) > \alpha, \\ -1 & \text{otherwise}. \end{cases}$$

Then, we obtain the following proposition (its proof can be found in Appendix B.1).

Proposition 4. *Chow's rule in multiclass classification with rejection can be expressed as*

$$f^*(\boldsymbol{x}) = \begin{cases} ® & \max_y f^{*,y}_{1-c}(\boldsymbol{x}) = -1, \\ \arg\max_y f^{*,y}_{1-c}(\boldsymbol{x}) & \text{otherwise}. \end{cases}$$

Proposition 4 suggests that by learning cost-sensitive classifiers $f^{*,y}_{1-c}$ for $y \in \mathcal{Y}$, it is possible to obtain Chow's rule without estimating the class-posterior probabilities. Note that when $c < 0.5$, there exists at most one $y' \in \mathcal{Y}$ such that $f^{*,y'}_{1-c}(\boldsymbol{x}) = 1$. This is because it implies that $\eta_{y'}(x) > 1-c$, which is larger than 0.5.

Table 1. Classification-calibrated binary surrogate losses and their properties including the convexity, symmetricity (i.e., $\phi(z) + \phi(-z)$ is a constant), and its capability to estimate the class posterior probability $\eta_1(\boldsymbol{x})$ in binary classification. The column "confidence-based" indicates that a loss is applicable to the confidence-based approach and it satisfies all conditions required in order to use the previous work to derive its excess risk bound (Yuan & Wegkamp, 2010; Ni et al., 2019). On the other hand, our cost-sensitive approach can guarantee the existence of the excess risk bound as long as a loss ϕ is classification-calibrated (Bartlett et al., 2006).

Loss name	$\phi(z)$	Convex	Symmetric	Estimating $\eta_1(\boldsymbol{x})$	Confidence-based
Squared	$(1-z)^2$	✓	×	✓	✓
Squared hinge	$\max(0, 1-z)^2$	✓	×	✓	✓
Exponential	$\exp(-z)$	✓	×	✓	✓
Logistic	$\log(1+\exp(-z))$	✓	×	✓	✓
Hinge	$\max(0, 1-z)$	✓	×	×	×
Savage	$\left[(1+\exp(2z))^2\right]^{-1}$	×	×	✓	×
Tangent	$(2\arctan(z)-1)^2$	×	×	✓	×
Ramp	$\max(0, \min(1, 0.5-0.5z))$	×	✓	×	×
Sigmoid	$[1+\exp(z)]^{-1}$	×	✓	×	×

4. A Surrogate Loss for the Cost-sensitive Approach

In this section, we propose a surrogate loss for the cost-sensitive approach for classification with rejection.

It is known that given training data, directly minimizing the empirical risk with respect to ℓ_{01c} is computationally infeasible (Bartlett & Wegkamp, 2008; Ramaswamy et al., 2018). Therefore, many surrogate losses have been proposed to learn a classifier with rejection in practice (Yuan & Wegkamp, 2010; Cortes et al., 2016b; Ni et al., 2019). Here, we propose the cost-sensitive surrogate loss for classification with rejection. Let $\boldsymbol{g}(\boldsymbol{x}) = [g_1(\boldsymbol{x}), \dots, g_K(\boldsymbol{x})]^\top$, where $g_y(x)\colon \mathcal{X} \to \mathbb{R}$ is the score function for a class y. Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be a binary margin surrogate loss. A margin loss is a class of loss functions for binary classification that takes only one real-valued argument (Bartlett et al., 2006; Reid & Williamson, 2010). Table 1 illustrates examples of binary margin surrogate losses. With a choice of ϕ, we can define our proposed cost-sensitive surrogate loss as follows.

Definition 5. Given a binary margin surrogate loss ϕ and a pre-defined rejection cost c, *the cost-sensitive surrogate loss* for classification with rejection is defined as

$$\mathcal{L}_{\mathrm{CS}}^{c,\phi}(\boldsymbol{g}; \boldsymbol{x}, y) = c\phi\big(g_y(\boldsymbol{x})\big) + (1-c)\sum_{y' \neq y} \phi\big(-g_{y'}(\boldsymbol{x})\big).$$

Next, following the empirical risk minimization framework (Vapnik, 1998), a learning objective function can be straightforwardly obtained as follows:

$$\hat{R}^{\mathcal{L}_{\mathrm{CS}}^{c,\phi}}(\boldsymbol{g}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_{\mathrm{CS}}^{c,\phi}(\boldsymbol{g}; \boldsymbol{x}_i, y_i). \tag{6}$$

Note that regularization can also be applied in practice to avoid overfitting. Moreover, we want to emphasize that although it is theoretically suggested to learn an ensemble of classifiers to solve classification with rejection, in practice, by using linear-in-parameter models or neural networks with K-dimensional vectorial outputs, we can conveniently learn all K cost-sensitive binary classifiers together at once, which is $\boldsymbol{g}$.

After learning $\boldsymbol{g}$ by minimizing Eq. (6), we have to design how to reject $\boldsymbol{x}$. Following the optimal rejection rule in our Proposition 4, i.e., $\max_y f_{1-c}^{*,y}(\boldsymbol{x}) = -1$, we can directly obtain the following rejection rule:

$$\max_y g_y(\boldsymbol{x}) \leq 0. \tag{7}$$

Intuitively, Cond. (7) suggests to reject $\boldsymbol{x}$ if all $g_y(\boldsymbol{x})$ have low prediction confidence. One may interpret this type of rejection as *distance rejection* (Dubuisson & Masson, 1993), where the rejection is made when $\boldsymbol{g}$ is uncertain whether $\boldsymbol{x}$ belongs to any of the known classes. Note that this does not necessarily imply that $\boldsymbol{x}$ belongs to an unknown class, e.g., $\boldsymbol{x}$ may be located close to the decision boundary, causing none of $g_y(\boldsymbol{x})$ to be confident enough to predict a class y.

Next, we also consider the following rejection rule:

$$\exists y, y' \text{ s.t. } y \neq y' \text{and } g_y(\boldsymbol{x}), g_{y'}(\boldsymbol{x}) > 0. \tag{8}$$

Cond. (8) suggests to reject $\boldsymbol{x}$ because there exists a prediction conflict among at least two binary classifiers, i.e., $g_y(\boldsymbol{x})$ suggests to predict a class y but $g_{y'}(\boldsymbol{x})$ suggests to predict another class y'. Note that if we succeed to obtain the optimal classifier $\boldsymbol{g}^*$, this condition is *impossible* to be satisfied. Recall that in Section 3.4, for $\boldsymbol{g}^*$, at most one $g_y^*(\boldsymbol{x})$ can be more than zero since it implies $\eta_y > 1 - c > 0.5$. Nevertheless, Cond. (8) may hold in practice due to empirical estimation. This rejection condition can be interpreted as *ambiguity rejection* (Dubuisson & Masson, 1993), where the rejection is made when $\boldsymbol{g}$ interprets $\boldsymbol{x}$ to be associated

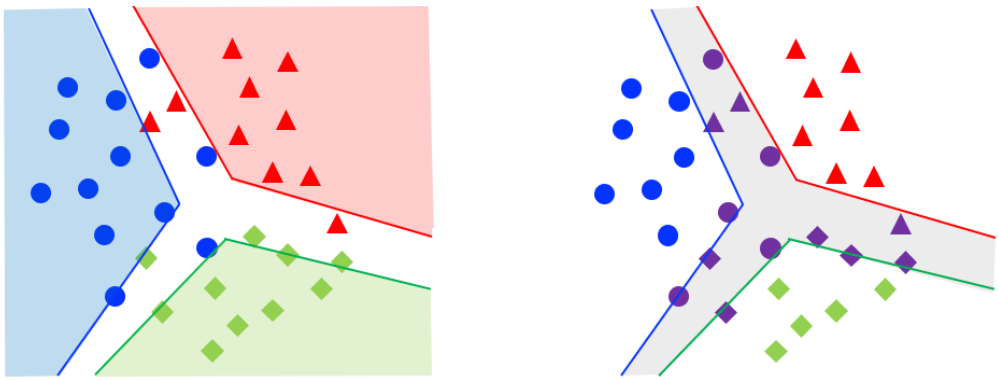

Figure 3. Illustration of the cost-sensitive approach. Figure (a) denotes a prediction function. Unlike the confidence-based approach (Figure 1), the prediction function is not designed to predict all data points in the space and the rejection region does not spread from the decision boundary. The decision boundary is based on an ensemble of cost-sensitive classifiers for blue, red, and green classes, respectively. Then, the rejector in figure (b) is constructed based on the rejection rule in Eq. (7) by aggregating the prediction result of each cost-sensitive classifier. Data points in purple are rejected.

with more than one class. More discussion on the proposed rejection conditions is provided in Appendix. D.

To sum up, we employ the following classification rule for the cost-sensitive approach:

$$f(\boldsymbol{x}; \boldsymbol{g}) = \begin{cases} \text{®} & \text{Conds. (7) or (8),} \\ \arg\max_y g_y(\boldsymbol{x}) & \text{otherwise.} \end{cases} \tag{9}$$

Figure 3 illustrates the cost-sensitive approach. It is worth mentioning that our rejection condition is different from that of Zhang et al. (2018). In their rejection rule, an input $\boldsymbol{x}$ is rejected if all binary classifiers' outputs are close to zero. In our case, Cond. (7) rejects $\boldsymbol{x}$ as long as all $g_y(\boldsymbol{x})$'s are negative, e.g., $\boldsymbol{x}$ is also rejected if all prediction outputs are much smaller than zero. Also, their method can predict a set of labels when at least two classifiers predict positively, which is different from our problem setting, where it is only allowed to predict one label or refrain from making a prediction.

5. Theoretical Analysis

In this section, we show that the classification rule $f(\boldsymbol{x}; \boldsymbol{g})$ in Eq. (9) can achieve the Chow's rule and also provide excess risk bounds.

5.1. Calibration

We begin by introducing the well-known notion of classification-calibrated loss in binary classification. Let us define the pointwise conditional surrogate risk for a fixed input $\boldsymbol{x}$ in binary classification with its class-posterior probability of a positive class η_1:

$$C^{\phi}_{\eta_1}(v) = \eta_1 \phi(v) + (1 - \eta_1)\phi(-v), \tag{10}$$

for $v \in \mathbb{R}$. Next, a classification-calibrated loss can be defined as follows.

Definition 6 (Bartlett et al. (2006)). *We say a loss ϕ is classification-calibrated if for any $\eta_1 \neq \frac{1}{2}$, we have*

$$\inf_{v(2\eta_1 - 1) \leq 0} C^{\phi}_{\eta_1}(v) > \inf_{v} C^{\phi}_{\eta_1}(v).$$

Intuitively, classification-calibration ensures that minimizing a loss ϕ can give the Bayes optimal binary classifier $\mathrm{sign}(2\eta_1 - 1)$. It is known that a convex loss ϕ is classification-calibrated if and only if it is differentiable at 0 and $\phi'(0) < 0$ (Bartlett et al., 2006).

Analogously, in classification with rejection, calibration is also an important property that has been used to verify if a surrogate loss is appropriate (Bartlett & Wegkamp, 2008; Yuan & Wegkamp, 2010; Cortes et al., 2016b;a; Ni et al., 2019). Calibration guarantees that by replacing the zero-one-c loss ℓ_{01c} with a surrogate loss $\mathcal{L}$, the optimal Chow's rule can still be obtained by minimizing the surrogate risk. Verifying the calibration condition in classification with rejection has not been as well-studied as ordinary binary classification. We are only aware of the works by Yuan & Wegkamp (2010) and Ni et al. (2019), which provided a condition to verify calibration of general loss functions for the confidence-based approach. Nevertheless, their condition can only verify a convex loss. Note that losses that are calibrated in our cost-sensitive approach may not be calibrated in the confidence-based approach, e.g., the sigmoid loss. See Table 1 for more details.

Now we are ready to show that the calibration condition of our proposed approach is equivalent to the classification-calibration condition of ϕ. Let us define the pointwise conditional surrogate risk $W_{\mathcal{L}}$ of an input $\boldsymbol{x}$ with its class-posterior probability $\boldsymbol{\eta}(\boldsymbol{x})$ for the multiclass case:

$$W_{\mathcal{L}}\big(\boldsymbol{g}(\boldsymbol{x}));\boldsymbol{\eta}(\boldsymbol{x})\big) = \sum_{y\in\mathcal{Y}} \eta_y(\boldsymbol{x})\mathcal{L}\big(\boldsymbol{g};\boldsymbol{x},y\big). \tag{11}$$

By analyzing the classification rule with respect to the conditional risk minimizer, we obtain the following theorem (its proof can be found in Appendix B.2).

Theorem 7. *Let g^* be a conditional risk minimizer that minimizes the pointwise conditional surrogate risk $g^*(\boldsymbol{x}) = \arg\min_g W_{\mathcal{L}_{\mathrm{CS}}^{c,\phi}}\big(\boldsymbol{g}(\boldsymbol{x});\boldsymbol{\eta}(\boldsymbol{x}))\big)$. The surrogate loss $\mathcal{L}_{\mathrm{CS}}^{c,\phi}$ is calibrated for classification with rejection, that is, $f(\boldsymbol{x};\boldsymbol{g}^*) = f^*(\boldsymbol{x})$ for all $\boldsymbol{x}\in\mathcal{X}$, if and only if ϕ is classification-calibrated.*

Theorem 7 suggests that the condition to verify if our cost-sensitive surrogate loss $\mathcal{L}_{\mathrm{CS}}^{c,\phi}$ is calibrated is equivalent to the condition of whether ϕ is classification-calibrated. As long as a binary surrogate loss ϕ is classification-calibrated, minimizing the surrogate risk w.r.t. $\mathcal{L}_{\mathrm{CS}}^{c,\phi}$ can lead to the optimal Chow's rule. As a result, Theorem 7 successfully borrows the knowledge in the literature of binary classification to help prove calibration in classification with rejection for the cost-sensitive approach.

5.2. Excess Risk Bound

While calibration ensures that the optimal solution w.r.t. a surrogate loss agrees with the optimal Chow's rule, an excess risk bound provides a regret bound relationship between the zero-one-c loss ℓ_{01c} and a surrogate loss $\mathcal{L}$.

Let $R_{1-c}^{\phi,i}(g_i)$ be the cost-sensitive binary surrogate risk for class i and $R^{\mathcal{L},*}$ be the minimum risk w.r.t. to the loss $\mathcal{L}$. We prove the following theorem, which is our main result to use for deriving the excess risk bound of the cost-sensitive approach for any classification-calibrated loss (its proof can be found in Appendix B.3).

Theorem 8. *Consider a cost-sensitive surrogate loss $\mathcal{L}_{\mathrm{CS}}^{c,\phi}$. Let f be a classification rule of the cost-sensitive approach with respect to the score function $\boldsymbol{g}$, that is, $f = f(\boldsymbol{x};\boldsymbol{g})$ for an input $\boldsymbol{x}$. If a binary surrogate loss ϕ is classification-calibrated, the excess risk bound can be expressed as follows:*

$$R^{\ell_{01c}}(f) - R^{\ell_{01c},*} \le R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}}(\boldsymbol{g}) - R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}},*} \tag{12}$$

$$\le \sum_{i=1}^{K} \psi_{\phi,1-c}^{-1}(R_{1-c}^{\phi,i}(g_i) - R_{1-c}^{\phi,i,*}), \tag{13}$$

where $\psi_{\phi,1-c}\colon \mathbb{R}\to\mathbb{R}$ is a non-decreasing invertible function and $\psi_{\phi,1-c}(0) = 0$.

Ineq. (12) suggests that the regret of the classification with rejection problem can be bounded by the regret of the cost-sensitive surrogate with respect to the zero-one loss ℓ_{01}. This inequality allows us to borrow the existing findings of cost-sensitive classification to give excess risk bounds for classification with rejection. Next, Ineq. (13) suggests that $R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}}(\boldsymbol{g}) - R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}},*}$ is bounded by the sum of an invertible function of the regret of the cost-sensitive binary classification risk. The invertible function $\psi_{\phi,1-c}$ is a well-studied function in the literature of cost-sensitive classification, which is guaranteed to exist for any classification-calibrated loss (Steinwart, 2007; Scott, 2012). For example, $\psi_{\phi,1-c}^{-1}(\epsilon) = \frac{\epsilon^2}{2c(1-c)-(\epsilon)(1-2c)}$ for the squared loss, where $\epsilon \ge 0$. Examples of $\psi_{\phi,1-c}$ for more losses and how to derive $\psi_{\phi,1-c}$ can be found in Steinwart (2007) and Scott (2012). Since $\psi_{\phi,1-c}$ is non-decreasing and $\psi_{\phi,1-c}(0) = 0$, the regret with respect to the zero-one-c loss will also get smaller and eventually become zero if the surrogate risk is successfully minimized.

As an example to demonstrate how to obtain an excess risk bound with our Theorem 8, we prove that the following excess risk bound holds for the hinge loss ϕ_{hin}, which is the loss that cannot estimate the class-posterior probabilities (Cortes & Vapnik, 1995), and its optimal solution for the confidence-based approach cannot mimic Chow's rule. The bound can be straightforwardly derived based on our Theorem 8 and the known fact that $\psi_{\phi_{\mathrm{hin}},1-c}^{-1}(\epsilon) = \epsilon$ (Steinwart, 2007).

Corollary 9. *Let us consider the hinge loss $\phi_{\mathrm{hin}}(z) = \max(0, 1-z)$. The excess risk bound for the cost-sensitive surrogate $\mathcal{L}_{\mathrm{CS}}^{c,\phi_{\mathrm{hin}}}$ can be given as follows:*

$$R^{\ell_{01c}}(f) - R^{\ell_{01c},*} \le R^{\mathcal{L}_{\mathrm{CS}}^{c,\phi_{\mathrm{hin}}}}(\boldsymbol{g}) - R^{\mathcal{L}_{\mathrm{CS}}^{c,\phi_{\mathrm{hin}}},*}.$$

6. Experimental Results

In this section, we provide experimental results of classification with rejection. The evaluation metric is the test empirical zero-one-c risk over ten trials. We also reported the rejection ratio, accuracy of accepted data, and the full experimental results in the table format in Appendix D. The varying rejection costs ranged from $\{0.1, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40\}$ for all settings. For noisy-labeled classification, we used the uniform noise (Angluin & Laird, 1988; Ghosh et al., 2015), where the randomly selected 25% of the training labels were flipped.

6.1. Experiment Setup

Datasets and models: For binary classification, we used the subjective-versus-objective classification (Subj), which is a text dataset (Pang & Lee, 2004). Moreover, we used Phishing and Spambase, which are tabular datasets, and

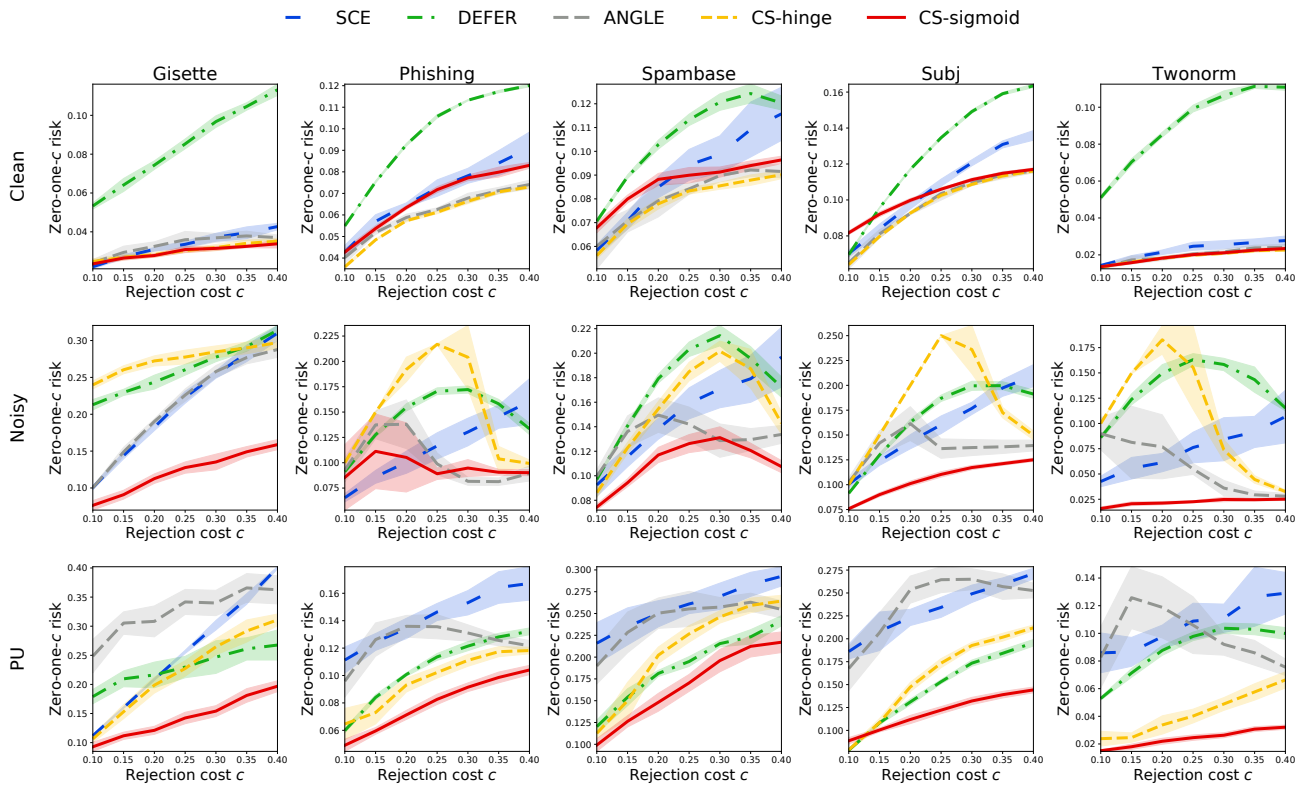

Figure 4. Mean and standard error of the test empirical zero-one-c risk over ten trials with varying rejection costs (Binary classification). Each column indicates the performance with respect to one dataset. (Top) clean-labeled classification with rejection. (Middle) noisy-labeled classification with rejection. (Bottom) PU-classification with rejection.

Twonorm, which is a synthetic dataset drawn from different multivariate Gaussian distributions (Lichman et al., 2013). We also used the Gisette dataset, which is the problem of separating the highly confusible digits 4 and 9 with noisy features (Guyon et al., 2005). Linear-in-input models were used for all binary datasets. For multiclass classification, we used Gas-Drift (Vergara et al., 2012) and Human activity recognition (HAR) (Anguita et al., 2013), which are tabular datasets. Multilayer perceptrons were used for both datasets. We also used the image datasets, which are MNIST (LeCun, 1998), Kuzushiji-MNIST (KMNIST) (Clanuwat et al., 2018), and Fashion-MNIST (Xiao et al., 2017). Convolutional neural networks were used for all image datasets. The implementation was done using PyTorch (Paszke et al., 2019). More details on the datasets and implementation can be found in Appendix C.

Methods: For the confidence-based approach, based on Ni et al. (2019), we used the softmax cross-entropy loss (SCE). For the classifier-reject approach, we used the proposed method by Mozannar & Sontag (2020) (DEFER). We also used the method by Zhang et al. (2018) with the bent hinge loss (ANGLE). For our cost-sensitive approach, we used the hinge (CS-hinge), and sigmoid (CS-sigmoid) losses.

Hyperparameter tuning: We provided additional training data for SCE and ANGLE to tune their hyperparameters. For SCE, we also added temperature scaling (Guo et al., 2017) to improve the prediction confidence. For ANGLE, we chose the bending slope paramater according to Zhang et al. (2018) and tuned the rejection threshold. In PU-classification, it is difficult to tune hyperparameters for them. Thus, we provided clean-labeled data for them *only* for hyperparameter tuning. Both rejection threshold of ANGLE and the temperature parameter for SCE are chosen from the following candidate set of twenty numbers spaced evenly in a log scale from 0 to 1 (inclusively) and nine integers from 2 to 10. Since only SCE and ANGLE require additional data to tune hyperparameters, it is not straightforward to provide a fair comparison because our methods and DEFER do not use validation data. Nevertheless, with less data, our methods are still competitive and can outperform the baselines in several settings.

6.2. Binary Classification with Rejection

Here, we compare the performance of all methods in clean-labeled, noisy-labeled, and positive-unlabeled classification with rejection. For PU-classification, we implemented all

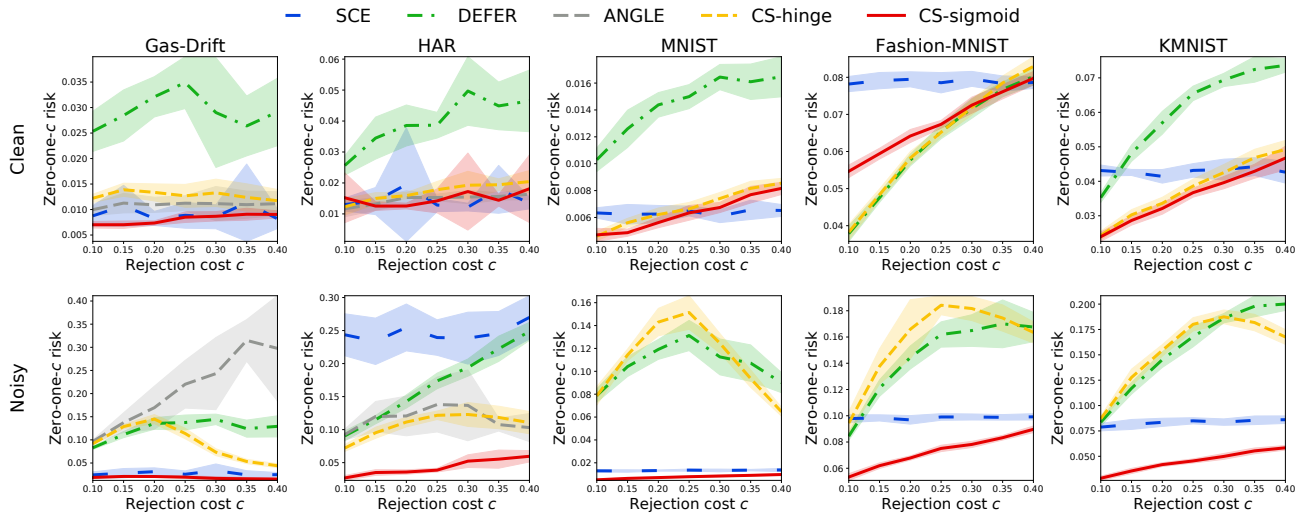

Figure 5. Mean and standard error of the test empirical zero-one-c risk over ten trials with varying rejection cost (Multiclass classification). Each column indicates the performance with respect to one dataset. (Top) clean-labeled classification with rejection. (Bottom) noisy-labeled classification with rejection. For MNIST, Fashion-MNIST, and KMNIST, we found that ANGLE failed miserably and has zero-one-c risk more than 0.5 and thus it is excluded from the figure for readability.

methods based on the empirical risk minimization framework proposed by Kiryo et al. (2017) (more detail can be found in Appendix C).

Figure 4 shows the performance with varying rejection costs for all settings. In clean-labeled classification, it can be observed that CS-hinge and ANGLE are most preferable in this setting. In noisy-labeled and PU classification, CS-sigmoid outperformed other methods in most cases. This illustrates the usefulness of having a flexible choice of loss functions. We also found that noise can degrade the performance to be worse than always reject for some methods. Moreover, we found that DEFER rejected data more often than other methods, which may sometimes lead to worse performance. In PU-classification, SCE and ANGLE did not perform well although clean labeled data were used for hyperparameter tuning. This could be due to a steep loss can suffer severely from the negative risk problem (Kiryo et al., 2017), causing them to be ineffective in PU-classification.

6.3. Multiclass Classification with Rejection

Figure 5 illustrates the performance of all methods in the clean-labeled and noisy-labeled settings. It can be observed that SCE had almost the same performance for all rejection costs. Although temperature scaling is applied, it seems that SCE still suffered from overconfidence (Guo et al., 2017) and failed to reject the ambiguous data points. This could be due to SCE has the high accuracy on the validation set (more than 90%) and thus temperature scaling could not smoothen the prediction confidence to reject the ambiguous data effectively. Interestingly, DEFER did not suffer from such overconfidence although it is also based on the cross-entropy loss and it rejected the data more than other methods. For ANGLE, we found that although it can perform competitively in Gas-drift and HAR, it failed miserably in the image datasets. For figure's readability, we report the performance of ANGLE in a table format in Appendix D. In noisy-labeled classification, CS-sigmoid outperformed other methods in most cases.

7. Conclusions

We have proposed a cost-sensitive approach to classification with rejection, where any classification-calibrated loss can be applied with theoretical guarantee. Our theory of excess risk bounds explicitly connects the classification with rejection problem to the cost-sensitive classification problem. Our experimental results using clean-labeled, noisy-labeled, and positive and unlabeled training data demonstrated the advantages of avoiding class-posterior probability estimation and having a flexible choice of loss functions.

Acknowledgements

We would like to thank Han Bao, Takeshi Teshima, Chenri Ni, and Junya Honda for helpful discussion, and also the Supercomputing Division, Information Technology Center, The University of Tokyo, for providing us the Reedbush supercomputer system to conduct the experiments. NC was supported by MEXT scholarship, JST AIP Challenge, and Google PhD Fellowship program. ZC was supported by JST AIP Challenge. MS was supported by the International Research Center for Neurointelligence (WPI-IRCN) at The University of Tokyo Institutes for Advanced Study.

References

Angluin, D. and Laird, P. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

Anguita, D., Ghio, A., Oneto, L., Parra, X., and Reyes-Ortiz, J. L. A public domain dataset for human activity recognition using smartphones. In *Esann*, volume 3, pp. 3, 2013.

Bao, H. and Sugiyama, M. Calibrated surrogate maximization of linear-fractional utility in binary classification. *AISTATS*, 2020.

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

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *JASA*, 101(473):138–156, 2006.

Begg, M. D. and Lagakos, S. On the consequences of model misspecification in logistic regression. *Environmental health perspectives*, 87:69–75, 1990.

Bickel, S., Brückner, M., and Scheffer, T. Discriminative learning for differing training and test distributions. In *ICML*, pp. 81–88, 2007.

Charoenphakdee, N. and Sugiyama, M. Positive-unlabeled classification under class prior shift and asymmetric error. In *SDM*, pp. 271–279. SIAM, 2019.

Charoenphakdee, N., Lee, J., and Sugiyama, M. On symmetric losses for learning from corrupted labels. *ICML*, 2019.

Chow, C. K. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, EC-6(4):247–254, 1957.

Chow, C. K. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16 (1):41–46, 1970.

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, D. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Cortes, C., DeSalvo, G., and Mohri, M. Boosting with abstention. In *Advances in Neural Information Processing Systems*, pp. 1660–1668, 2016a.

Cortes, C., DeSalvo, G., and Mohri, M. Learning with rejection. In *ALT*, pp. 67–82, 2016b.

Dembczynski, K., Jachnik, A., Kotlowski, W., Waegeman, W., and Hüllermeier, E. Optimizing the f-measure in multi-label classification: Plug-in rule approach versus structured loss minimization. In *ICML*, pp. 1130–1138, 2013.

du Plessis, M. C., Niu, G., and Sugiyama, M. Analysis of learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems*, pp. 703–711, 2014.

du Plessis, M. C., Niu, G., and Sugiyama, M. Convex formulation for learning from positive and unlabeled data. In *ICML*, pp. 1386–1394, 2015.

Dubuisson, B. and Masson, M. A statistical decision rule with incomplete knowledge about classes. *Pattern recognition*, 26(1):155–165, 1993.

El-Yaniv, R. and Wiener, Y. On the foundations of noise-free selective classification. *JMLR*, 11(May):1605–1641, 2010.

Elkan, C. The foundations of cost-sensitive learning. In *IJCAI*, volume 17, pp. 973–978. Lawrence Erlbaum Associates Ltd, 2001.

Franc, V. and Prusa, D. On discriminative learning of prediction uncertainty. In *ICML*, pp. 1963–1971, 2019.

Geifman, Y. and El-Yaniv, R. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 4878–4887, 2017.

Ghosh, A., Manwani, N., and Sastry, P. Making risk minimization tolerant to label noise. *Neurocomputing*, 160: 93–107, 2015.

Grandvalet, Y., Rakotomamonjy, A., Keshet, J., and Canu, S. Support vector machines with a reject option. In *Advances in Neural Information Processing Systems*, pp. 537–544, 2009.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *ICML*, pp. 1321–1330. JMLR. org, 2017.

Guyon, I., Gunn, S., Ben-Hur, A., and Dror, G. Result analysis of the nips 2003 feature selection challenge. In *Advances in Neural Information Processing Systems*, pp. 545–552, 2005.

Heagerty, P. J. and Kurland, B. F. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, 88(4):973–985, 2001.

Hein, M., Andriushchenko, M., and Bitterwolf, J. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *CVPR*, pp. 41–50, 2019.

Herbei, R. and Wegkamp, M. H. Classification with reject option. *Canadian Journal of Statistics*, 34(4):709–721, 2006.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pp. 448–456. PMLR, 2015.

Kanamori, T., Suzuki, T., and Sugiyama, M. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86(3):335–367, 2012.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *ICLR*, 2015.

Kiryo, R., Niu, G., du Plessis, M. C., and Sugiyama, M. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems*, pp. 1674–1684, 2017.

Koyejo, O. O., Natarajan, N., Ravikumar, P. K., and Dhillon, I. S. Consistent binary classification with generalized performance metrics. In *Advances in Neural Information Processing Systems*, pp. 2744–2752, 2014.

LeCun, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Lichman, M. et al. UCI machine learning repository, 2013.

Manwani, N., Desai, K., Sasidharan, S., and Sundararajan, R. Double ramp loss based reject option classifier. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 151–163. Springer, 2015.

Mozannar, H. and Sontag, D. Consistent estimators for learning to defer to an expert. *ICML*, 2020.

Nair, V. and Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *ICML*, pp. 807–814, 2010.

Ni, C., Charoenphakdee, N., Honda, J., and Sugiyama, M. On the calibration of multiclass classification with rejection. In *Advances in Neural Information Processing Systems*, pp. 2582–2592, 2019.

Pang, B. and Lee, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058*, 2004.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pp. 8026–8037, 2019.

Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing*, pp. 1532–1543, 2014.

Qin, J. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998.

Ramaswamy, H. G., Tewari, A., Agarwal, S., et al. Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics*, 12(1): 530–554, 2018.

Reid, M. D. and Williamson, R. C. Composite binary losses. *JMLR*, 11:2387–2422, 2010.

Saerens, M., Latinne, P., and Decaestecker, C. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41, 2002.

Scott, C. Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics*, 6:958–992, 2012.

Steinwart, I. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, 2007.

Sugiyama, M., Suzuki, T., and Kanamori, T. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

Vapnik, V. *Statistical learning theory. 1998*, volume 3. Wiley, New York, 1998.

Vergara, A., Vembu, S., Ayhan, T., Ryan, M. A., Homer, M. L., and Huerta, R. Chemical gas sensor drift compensation using classifier ensembles. *Sensors and Actuators B: Chemical*, 166:320–329, 2012.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

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

Zhang, C., Wang, W., and Qiao, X. On reject and refine options in multicategory classification. *Journal of the American Statistical Association*, 113(522):730–745, 2018.

Zhang, T. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pp. 56–85, 2004.