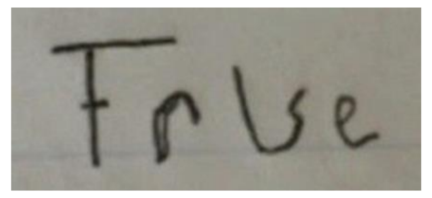
Classification with Rejection Based on Cost-sensitive Classification

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

[1] The University of Tokyo, [2] RIKEN AIP

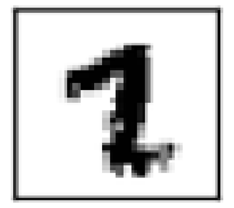
International conference on machine learning (ICML) 2021

Summary



True or False?

https://me.me/i/the-right-way-to-answer-true-and-false-questions-18781463



1,2 or 7?

MNIST dataset (Lecun, 1998)

Need surgery?

https://www.iconfinder.com/icons/5692300/ai_doctor_medical_nurse_robot_stethoscope_icon

Saying **"I don't know"** can **prevent misclassification**.
How to learn a classifier to say **"I don't know"** reasonably?

The well-known confidence-based approach typically requires estimating $p(y|\boldsymbol{x})$.
Theoretical framework typically requires a loss to be convex. (Ni et al., 2019)
Existing approaches have less loss choice than that of ordinary classification.

Contributions:

We propose a cost-sensitive approach to classification with rejection.

1. It can **avoid estimating** $p(y|\boldsymbol{x})$.
2. It is applicable to **both binary and multiclass cases**.
3. It is **theoretically justifiable for any classification-calibrated loss***.

*Classification calibration is known to be a minimum loss requirement for ordinary classification.

Problem formulation

Given: rejection cost $c \in (0, 0.5)$, training input-output pairs:

$$\{\boldsymbol{x}_i, y_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}, y)$$

$\mathcal{Y} = \{1, 2, \ldots, K\}$: Label space
$f : \mathcal{X} \to \mathcal{Y} \cup \{Ⓡ\}$: Classification rule

Goal: minimize 0-1-c risk:

$$R^{\ell_{01c}}(f) = \underset{(\boldsymbol{x},y)\sim p(\boldsymbol{x},y)}{\mathbb{E}} [\ell_{01c}(f(\boldsymbol{x}), y)]$$

$$\ell_{01c}(f(\boldsymbol{x}), y) = \begin{cases} c & \text{if } f(\boldsymbol{x}) = Ⓡ \quad \textbf{Rejection} \\ \underbrace{\ell_{01}(f(\boldsymbol{x}), y)}_{\textbf{0-1 loss}} & \text{otherwise} \quad \textbf{Prediction} \end{cases}$$

Rejection cost c is less than **misclassification cost**.

Directly minimizing the empirical 0-1-c risk is computationally infeasible. (Bartlett and Wegkamp, 2008)

Bayes-optimal solution: Chow's rule

Knowing $p(y|\boldsymbol{x})$ is sufficient to obtain optimal solution. (Chow 1970)

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

Straightforward solution: estimating $p(y|\boldsymbol{x})$ (confidence-based approach).

- ➢ **More restrictive loss requirement** than classification calibration. (Reid and Williamson, 2010)

Well-known loss such as hinge, ramp, and sigmoid losses are classification-calibrated but not capable of estimating $p(y|\boldsymbol{x})$.

Q: Can we have a framework that can use any classification-calibrated loss?

Proposal: Cost-sensitive approach

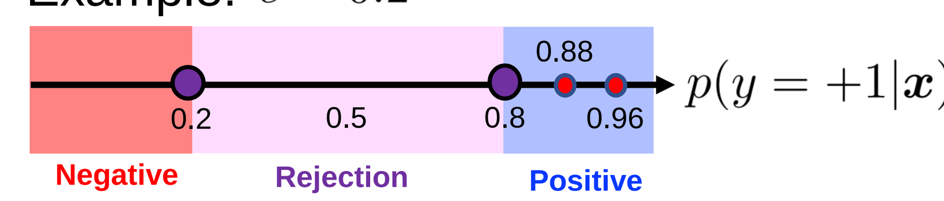
Binary case

Chow's rule for the binary case:
$$f^*(\boldsymbol{x}) = \begin{cases} +1 & p(y=1|\boldsymbol{x}) > 1 - c, \\ Ⓡ & c \le p(y=1|\boldsymbol{x}) \le 1 - c, \\ -1 & p(y=1|\boldsymbol{x}) < c, \end{cases}$$

To mimic Chow's rule, we only need to know:

1. $p(y=1|\boldsymbol{x}) > 1 - c$
2. $p(y=1|\boldsymbol{x}) < c$

Example: $c = 0.2$



Binary cost-sensitive classification:
Binary classification where false negative penalty $\neq$ false positive penalty.
Let false positive penalty be $\alpha \in (0, 1)$ and false negative penalty be $1 - \alpha$:

- ➢ Solving cost-sensitive classification can validate if $p(y=1|\boldsymbol{x}) > \alpha$.
- ➢ Loss requirement: ***classification calibration*** (Scott, 2012)

Proposal: cost-sensitive approach to binary classification with rejection.

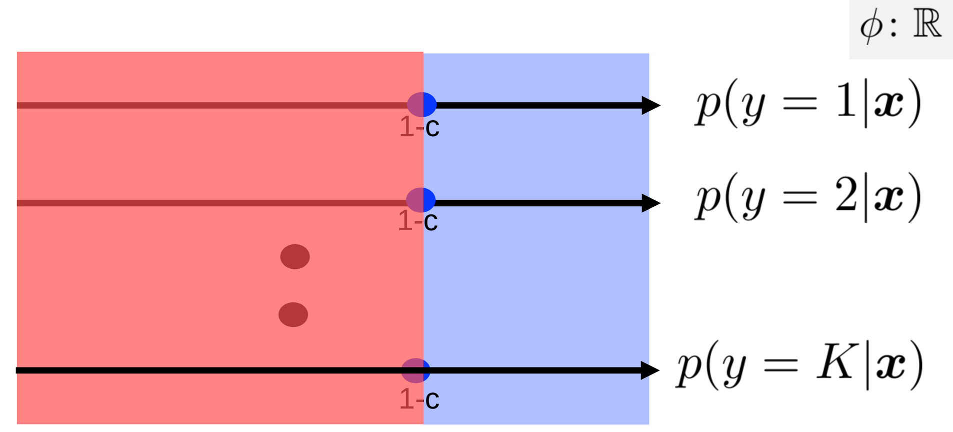
- ➢ Learn two cost-sensitive classifiers for $\alpha = c$ and $\alpha = 1 - c$.
- ➢ Predict if both classifiers predict the same class and reject otherwise.

Multiclass extension

Learn K **one-vs-rest cost-sensitive classifiers** with $\alpha = 1 - c$.
Can be learned at once by learning $\boldsymbol{g} : \mathcal{X} \to \mathbb{R}^K$.

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

$\phi : \mathbb{R} \to \mathbb{R}$: Classification-calibrated margin loss



Predict if:
Only one classifier returns positive

Reject if:
1. All classifiers predict negative, or
2. More than one classifier predicts positive

Classification rule:

$$f(\boldsymbol{x}; \boldsymbol{g}) = \begin{cases} Ⓡ & \boxed{\max_y g_y(\boldsymbol{x}) \le 0,} \\ Ⓡ & \boxed{\exists y, y' \text{ s.t. } y \neq y' \; g_y(\boldsymbol{x}), g_{y'}(\boldsymbol{x}) > 0,} \\ \arg\max_y g_y(\boldsymbol{x}) & \boxed{\text{otherwise.}} \end{cases}$$

Excess risk bound

Main result:
$$\underbrace{R^{\ell_{01c}}(f) - R^{\ell_{01c},*}}_{\textbf{Excess 0-1-c risk}} \le \underbrace{R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}}(\boldsymbol{g}) - R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}},*}}_{\textbf{Excess cost-sensitive 0-1 risk}}$$

Excess 0-1-c risk is bounded by excess cost-sensitive 0-1 risk!

Excess risk bound of cost-sensitive 0-1 risk is well studied. (Scott 2012, Steinwart, 2007)

$$\underbrace{R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}}}(\boldsymbol{g}) - R^{\mathcal{L}_{\mathrm{CS}}^{c,\ell_{01}},*} \le \sum_{i=1}^{K} \psi_{\phi,1-c}^{-1}\big(R_{1-c}^{\phi,i}(g_i) - R_{1-c}^{\phi,i,*}\big),}_{\textbf{Excess cost-sensitive surrogate risk}}$$

$\psi : \mathbb{R} \to \mathbb{R}$: Invertible increasing function
$\psi(0) = 0$

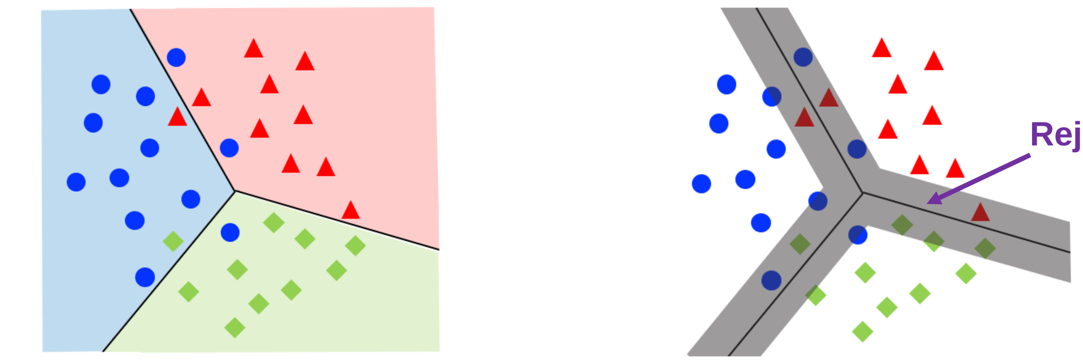
(please see our paper for more details.)

Excess 0-1-c risk is also bounded by excess cost-sensitive surrogate risk!

Connecting theory of cost-sensitive classification to classification with rejection!

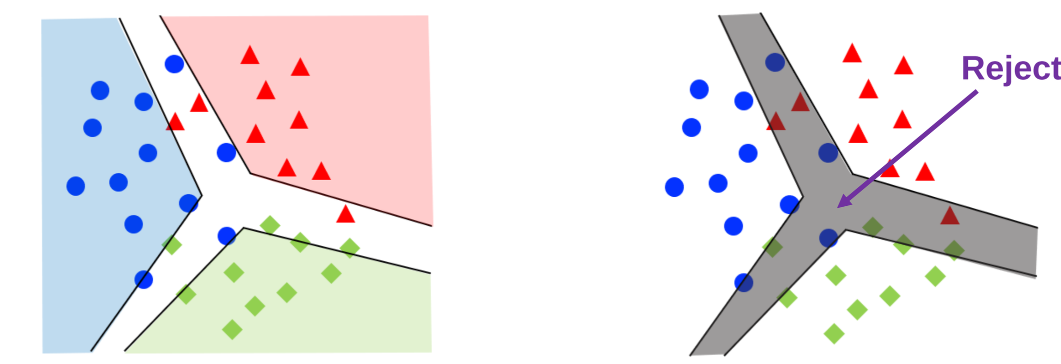
Comparison of approaches

Existing confidence-based approach



Rejection region spreads from the decision boundary.
Loss function choice is restrictive.

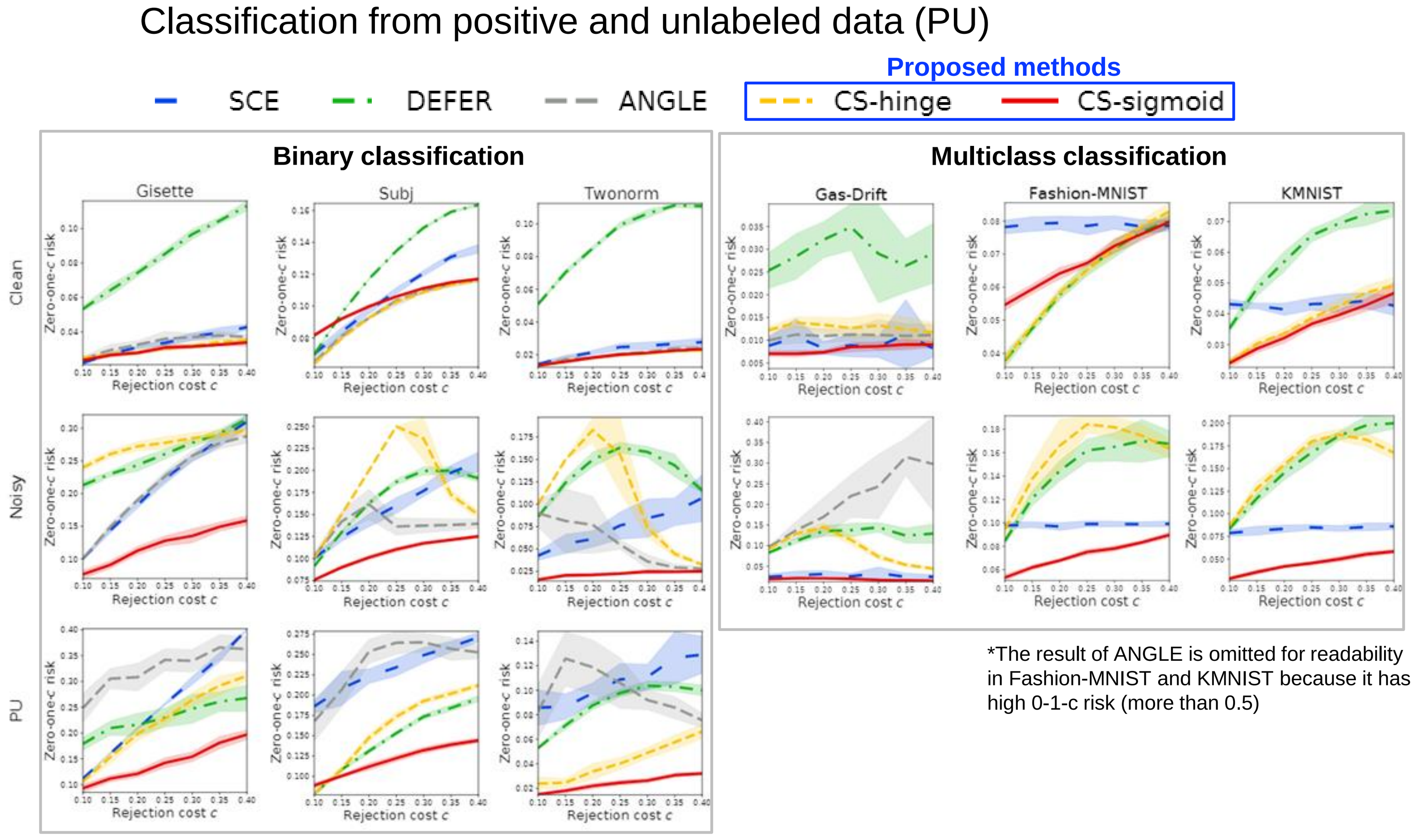
Proposed cost-sensitive approach



Rejection region is obtained by aggregating K-cost sensitive classifiers.
Loss requirement: ***classification calibration***

Experiments

Evaluation metric: Test empirical 0-1-c risk with varying rejection cost
Baseline: Softmax cross-entropy loss with temperature scaling (SCE), DEFER (Mozannar and Sontag, 2020), ANGLE (Zhang et al., 2017)
Setting: Clean-labeled classification (Clean), Noisy-labeled classification (Noisy), Classification from positive and unlabeled data (PU)



*The result of ANGLE is omitted for readability in Fashion-MNIST and KMNIST because it has high 0-1-c risk (more than 0.5)

CS-hinge works well in classification from clean labels (Clean).
CS-sigmoid works well in classification from noisy labels (Noisy) and classification from positive and unlabeled data (PU).
*sigmoid and hinge losses are classification-calibrated but not capable of estimating $p(y|\boldsymbol{x})$.

References

[1] LeCun, Y. The MNIST database of handwritten digits. http://yann. lecun. com/exdb/mnist/,1998.
[2] Ni, C., Charoenphakdee, N., Honda, J., and Sugiyama, M. On the calibration of multiclass classification with rejection. NeurIPS, 2019.
[3] Bartlett, P. L., and Wegkamp, M. H. Classification with a reject option using a hinge loss. JMLR, 2008.
[4] Chow., C. K. On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory, 16(1):41–46, 1970.
[5] Reid, M. D. and Williamson, R. C. Composite binary losses. JMLR, 11:2387–2422, 2010.
[6] Scott, C. Calibrated asymmetric surrogate losses. Electronic Journal of Statistics, 2012.
[7] Steinwart, I. How to compare different loss functions and their risks. Constructive Approximation, 2007.
[8] Mozannar, H., and Sontag, D. Consistent estimators for learning to defer to an expert, ICML, 2020
[9] Zhang, C., Wang, W., and Qiao, X. On reject and refine options in multicategory classification, JASA, 2017