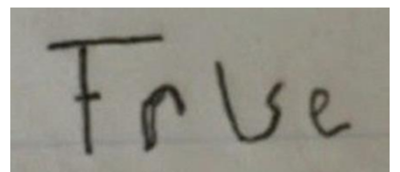
On the Calibration of Multiclass Classification with Rejection

Chenri Ni[1] Nontawat Charoenphakdee[1,2] Junya Honda[1,2] Masashi Sugiyama[2,1]

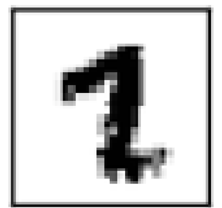
1: The University of Tokyo 2: RIKEN AIP

Introduction: Learning with rejection



True or False?

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



1,2 or 7?

Source: MNIST dataset
Lecun (1998)

Saying **"I don't know"** can **prevent misclassification**.

Related work:

Approach	Binary	Multiclass
Confidence-base	Bartlett+ (2008); Yuan+ (2010)	Ramaswamy+ (2018)
Classifier-rejector	Cortes+ (2015, 2016)	X

Ramaswamy+ (2018) ….

Contributions:

- **Calibration condition for surrogate losses in the classifier-rejector approach, which suggests the difficulty especially in the multiclass case**
- **Excess risk bounds and estimation error bounds to guarantee the one-vs-all (OVA) and cross-entropy (CE) losses in the confidence-based approach**

Multiclass classification with rejection

Chow (1970); Ramaswamy+ (2018)

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

Rejection cost: $c \in (0, 0.5)$

Find: Classifier: $f(\boldsymbol{x}) = \operatorname{argmax}_{y \in \mathcal{Y}} g_y(\boldsymbol{x}) \in \mathcal{Y}$

Rejector: $r(\boldsymbol{x}) \in \mathbb{R}$

$\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$

$y \in \mathcal{Y} = \{1, \ldots, K\}$

$g_i(\boldsymbol{x})\colon \mathcal{X} \to \mathbb{R}$

$$\mathrm{decision}(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}) & \text{if } r(\boldsymbol{x}) > 0 \\ \text{reject} & \text{otherwise} \end{cases}$$

Goal: Minimize $R_{0\text{-}1\text{-}c}(r, f) = \mathbb{E}_{p(\boldsymbol{x},y)}[\mathcal{L}_{0\text{-}1\text{-}c}(r, f; \boldsymbol{x}, y)]$

where $$\mathcal{L}_{0\text{-}1\text{-}c}(r, f; \boldsymbol{x}, y) = \underbrace{\mathbb{1}_{[f(\boldsymbol{x}) \neq y]} \mathbb{1}_{[r(\boldsymbol{x}) > 0]}}_{\text{misclassification loss}} + \underbrace{c \mathbb{1}_{[r(\boldsymbol{x}) \leq 0]}}_{\text{rejection loss}}$$

$\mathcal{L}_{0\text{-}1\text{-}c}(r, f; \boldsymbol{x}, y)$ **is difficult to directly optimize.**

Yuan+ (2010); Cortes+ (2015, 2016); Ramaswamy+ (2018)

A computationally-efficient and theoretically justified surrogate loss is needed.

Calibration

Calibration ensures that minimizing a surrogate loss will lead to an optimal solution.

Optimal solution of classification with rejection:

Chow (1970)

$$f^*(\boldsymbol{x}) = \arg\max_{y \in \mathcal{Y}} \eta_y(\boldsymbol{x})$$
$$r^*(\boldsymbol{x}) = \max_{y \in \mathcal{Y}} \eta_y(\boldsymbol{x}) - (1 - c)$$

$\eta_y(\boldsymbol{x}) = p(y|\boldsymbol{x})$

- (r, f) is **calibrated** if $R_{0\text{-}1\text{-}c}(r, f) = R_{0\text{-}1\text{-}c}(r^*, f^*)$.
- f is **classification-calibrated** if $f(\boldsymbol{x}) = f^*(\boldsymbol{x})$.
- r is **rejection-calibrated** if $\operatorname{sign}[r(\boldsymbol{x})] = \operatorname{sign}[r^*(\boldsymbol{x})]$.

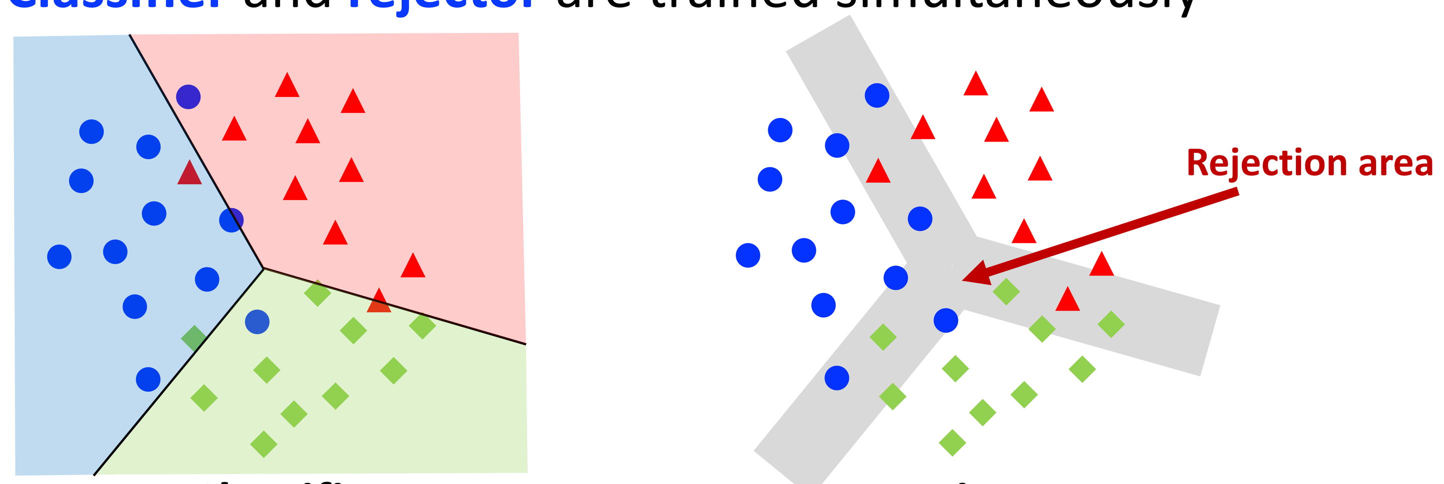
If (r, f) is calibrated, r must be rejection-calibrated.

A minimizer of a surrogate loss should give a calibrated (r, f).

Classifier-rejector approach

Cortes+ (2015, 2016)

Classifier and **rejector** are trained simultaneously



Classifier Rejector

Cortes+ (2015, 2016) proposed this approach in binary case:

- State-of-the-art method in binary case.
- Rejector is flexible, which is desirable when classifier model is misspecified.

$(r_{\boldsymbol{\eta}}^\dagger, f_{\boldsymbol{\eta}}^\dagger) = \operatorname*{arg\,min}_{r \in \mathbb{R},\, \boldsymbol{g} \in \mathbb{R}^K} W(r, f; \boldsymbol{\eta})$ $\quad \boldsymbol{\eta}(\boldsymbol{x}) = [\eta_1(\boldsymbol{x}), \ldots, \eta_K(\boldsymbol{x})]^\top$ $\quad W(r(\boldsymbol{x}), f(\boldsymbol{x}); \boldsymbol{\eta}(\boldsymbol{x})) = \sum_{y \in \mathcal{Y}} \eta_y(\boldsymbol{x}) \mathcal{L}(r, f; \boldsymbol{x}, y)$

Corollary 5: (Necessary condition for rejection calibration)

For $\mathcal{L}(r, f; \boldsymbol{x}, y)$ that is convex with respect to r and $\left.\frac{\partial^2 W(r, f_{\boldsymbol{\eta}}^\dagger; \boldsymbol{\eta})}{\partial r^2}\right|_{r=0} > 0$, $r^\dagger$ is rejection-calibrated only if both conditions hold:

Condition (1)

$$\underbrace{\sup_{\boldsymbol{\eta}: \max_y \eta_y = 1-c} \left.\frac{\partial W(r, f_{\boldsymbol{\eta}}^\dagger; \boldsymbol{\eta})}{\partial r}\right|_{r=0} = 0}_{\text{Condition for false reject rate to be zero}}$$

Condition (2)

$$\underbrace{\inf_{\boldsymbol{\eta}: \max_y \eta_y = 1-c} \left.\frac{\partial W(r, f_{\boldsymbol{\eta}}^\dagger; \boldsymbol{\eta})}{\partial r}\right|_{r=0} = 0}_{\text{Condition for false accept rate to be zero}}$$

A necessary and sufficient condition is also provided in our paper (Theorem 4)

Supremum and infimum values coincide under the same constraint.

When $\max_y \eta_y = 1 - c$

- Binary case: $\boldsymbol{\eta}$ can only be either $[1 - c, c]^\top$ or $[c, 1 - c]^\top$.
- Multiclass case: $\boldsymbol{\eta}$ **has infinitely many candidates!**

Case study:

$\alpha \in \mathbb{R}$ $\beta \in \mathbb{R}$ Hyperparameters
$\phi: \mathbb{R} \to \mathbb{R}$ $\psi: \mathbb{R} \to \mathbb{R}$ Convex margin losses

- **Multiplicative pairwise comparison (MPC) loss:**
$$\mathcal{L}_{\mathrm{MPC}}(r, f; \boldsymbol{x}, y) = \sum_{y' \neq y} \phi\Big(\alpha\big(g_y(\boldsymbol{x}) - g_{y'}(\boldsymbol{x})\big)\Big)\psi(-\alpha r(\boldsymbol{x})) + c\psi(\beta r(\boldsymbol{x}))$$
- **Additive pairwise comparison (APC) loss:**
$$\mathcal{L}_{\mathrm{APC}}(r, f; \boldsymbol{x}, y) = \sum_{y' \neq y} \phi\Big(\alpha\big(g_y(\boldsymbol{x}) - g_{y'}(\boldsymbol{x}) - r(\boldsymbol{x})\big)\Big) + c\psi(\beta r(\boldsymbol{x}))$$

Consider $\phi(z) = \psi(z) = \exp(-z)$

Condition (1) gives $\frac{\beta}{\alpha} = (K - 2) + 2\sqrt{(K-1)\frac{1-c}{c}}$

Condition (2) gives $\frac{\beta}{\alpha} = 2\sqrt{\frac{1-c}{c}}$

Equivalent to the result by Cortes+ (2016) when considering a binary case $(K = 2)$.

In multiclass case, (α, β) **satisfying both conditions does not exist.**

Similar results also hold when using the logistic loss $\phi(z) = \psi(z) = \log(1 + \exp(-z))$.

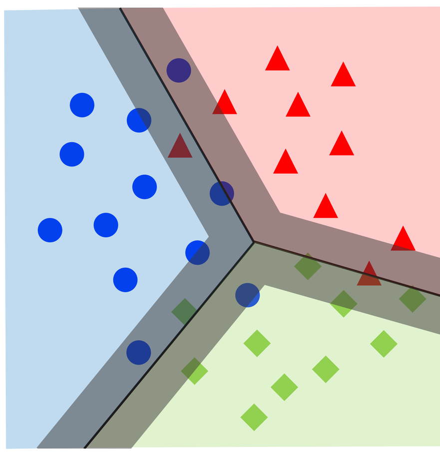
References

[1] C. K. Chow. On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory, 1970.
[2] P. L. Bartlett, M. H. Wegkamp. Classification with a reject option using a hinge loss. JMLR, 2008.
[3] C. Cortes, G. DeSalvo, and M. Mohri. Learning with rejection. ALT, 2015.
[4] C. Cortes G. DeSalvo, and M. Mohri. Boosting with abstention. NeurIPS, 2016.
[5] H.G. Ramaswamy, A. Tewari, and S. Agarwal. Consistent algorithms for multiclass classification with an abstain option. Electronic Journal of Statistics, 2018.
[6] M. Yuan, M.H. Wegkamp. Classification methods with reject option based on convex risk minimization. JMLR, 2010.
[7] Y. Lecun, The MNIST database of handwritten digits. http://yann.lecun.com/exdb/mnist/, 1998.

Confidence-based approach

Bartlett+ (2008); Yuan+ (2010); Ramaswamy+ (2018)

Rejector depends solely on **classifier's** confidence



- **Cross-entropy (CE) loss:**
$$\mathcal{L}_{\mathrm{CE}}(f; \boldsymbol{x}, y) = -g_y(\boldsymbol{x}) + \log \sum_{y' \in \mathcal{Y}} \exp\big(g_{y'}(\boldsymbol{x})\big)$$
- **One-versus-all (OVA) loss:**
$$\mathcal{L}_{\mathrm{OVA}}(f; \boldsymbol{x}, y) = \phi\big(g_y(\boldsymbol{x})\big) + \sum_{y' \neq y} \phi\big(-g_{y'}(\boldsymbol{x})\big)$$
- **Rejector:** $\boldsymbol{g}(\boldsymbol{x}) = [g_1(\boldsymbol{x}), \ldots, g_K(\boldsymbol{x})]^\top$
$$r_f(\boldsymbol{x}) = \max_{y \in \mathcal{Y}} \Psi^{-1}(\boldsymbol{g}(\boldsymbol{x})) - (1 - c)$$

$\Psi^{-1}: \mathbb{R}^K \to [0, 1]^K$ Inverse link function

$\Psi_{y,\mathrm{OVA}}^{-1}(\boldsymbol{g}) = \frac{\phi'(-g_y)}{\phi'(-g_y) + \phi'(g_y)}$

See our paper for conditions on ϕ.

$\Psi_{y,\mathrm{CE}}^{-1}(\boldsymbol{g}) = \frac{\exp(g_y)}{\sum_{y' \in \mathcal{Y}} \exp(g_{y'})}$

Softmax function

We provide excess risk bounds to guarantee OVA and CE losses.

Excess risk:

$$\Delta R_{0\text{-}1\text{-}c}(r_f, f) = R_{0\text{-}1\text{-}c}(r_f, f) - \inf_{f':\text{measurable}} R_{0\text{-}1\text{-}c}(r_f, f)$$
$$\Delta R_\ell(f) = R_\ell(f) - \inf_{f':\text{measurable}} R_\ell(f')$$

Excess risk bound of OVA loss:

$$(2C)^{-s} \Delta R_{0\text{-}1\text{-}c}(r_f, f)^s \leq \Delta R_{\mathrm{OVA}}(f)$$

Loss Name	$\phi(z)$	C	s
Logistic	$\log(1 + \exp(-z))$	$\frac{1}{2}$	2
Exponential	$\exp(-z)$	$\frac{1}{\sqrt{2}}$	2
Squared	$(1 - z)^2$	$\frac{1}{2}$	2
Squared Hinge	$(1 - z)_+^2$	$\frac{1}{2}$	2

Extension of the result by Yuan+ (2010) to the multiclass case.

Excess risk bound of CE loss:

$$\frac{1}{2} \Delta R_{0\text{-}1\text{-}c}(r_f, f)^2 \leq \Delta R_{\mathrm{CE}}(f)$$

Needs analysis specific to the multiclass case where previous techniques cannot be applied.

Minimizers of OVA and CE losses also minimize the 0-1-c loss.

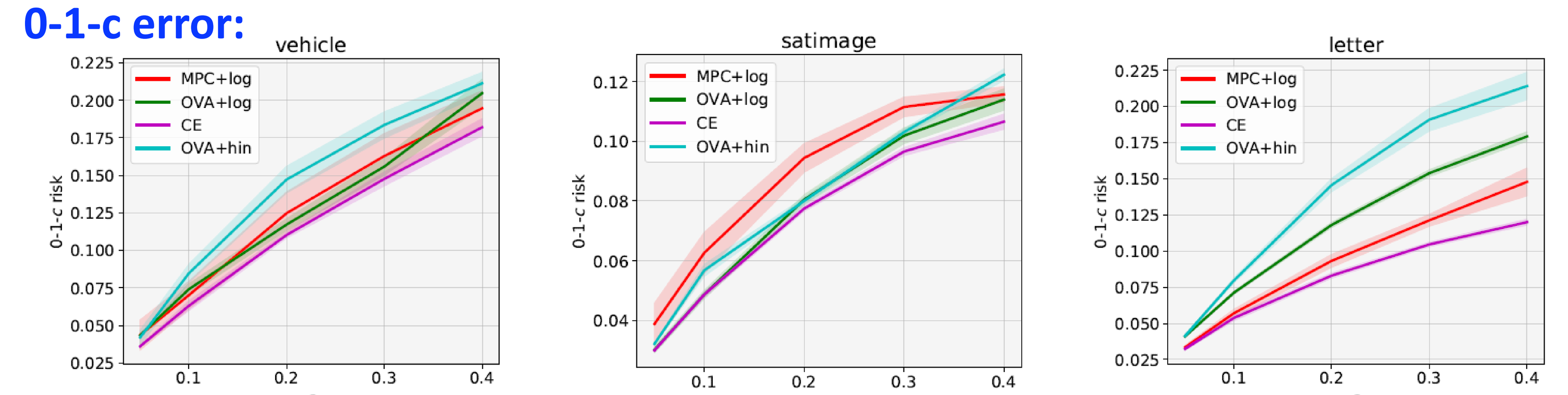
See our paper for estimation error bound using Rademacher complexity.

Experiments

Classifier-rejector: MPC+log (MPC with logistic loss), APC+log (APC with logistic loss)

Confidence-based: OVA+hin by Ramaswamy+ (2018), OVA+log (OVA with logistic loss), CE

0-1-c error:



Accuracy of non-rejected data: "- (-)" indicates all data were rejected.

dataset	c	APC+log	MPC+log	OVA+log	CE
vehicle	0.05	- (-)	96.6 (2.3)	100 (0.0)	**100 (0.0)**
	0.2	98.4 (1.9)	92.4 (3.0)	97.9 (0.7)	**97.4 (0.1)**
	0.4	**89.1 (2.9)**	85.3 (4.2)	90.2 (1.6)	**91.7 (0.9)**
satimage	0.05	**99.1 (0.2)**	97.2 (1.4)	**98.7 (0.1)**	**98.3 (0.1)**
	0.2	95.0 (1.0)	92.6 (1.2)	96.2 (0.2)	**95.7 (0.1)**
	0.4	91.5 (0.7)	89.0 (1.1)	92.2 (0.3)	**91.8 (0.2)**
yeast	0.05	- (-)	- (-)	- (-)	- (-)
	0.2	- (-)	- (-)	- (-)	**80.6 (6.2)**
	0.4	- (-)	- (-)	75.0 (3.9)	**76.6 (1.7)**

dataset	c	APC+log	MPC+log	OVA+log	CE
covtype	0.05	**79.5 (2.1)**	79.8 (1.7)	**82.1 (2.7)**	**82.0 (3.2)**
	0.2	74.0 (1.8)	73.8 (1.0)	74.9 (1.4)	**77.1 (0.3)**
	0.4	**69.8 (1.3)**	64.9 (3.4)	**68.7 (1.1)**	**69.4 (1.8)**
letter	0.05	**99.8 (0.1)**	98.6 (0.2)	**99.6 (0.2)**	**99 8 (0.0)**
	0.2	97.9 (0.3)	96.9 (0.5)	**98.3 (0.2)**	**98.4 (0.1)**
	0.4	**95.2 (0.5)**	94.6 (3.8)	94.6 (0.2)	**94.9 (0.3)**