

On the Calibration of Multiclass Classification with Rejection

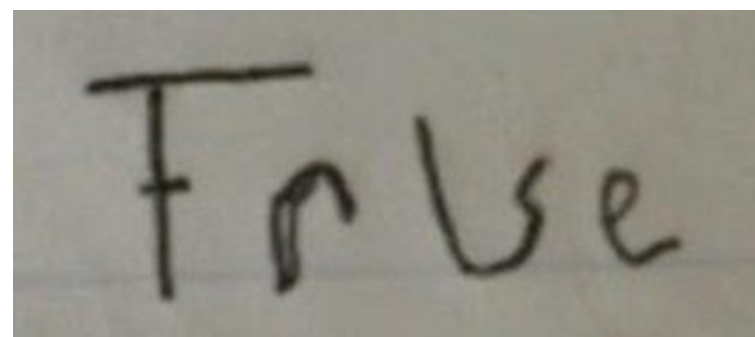
Chenri Ni¹ Nontawat Charoenphakdee^{1,2} Junya Honda^{1,2} Masashi Sugiyama^{2,1}



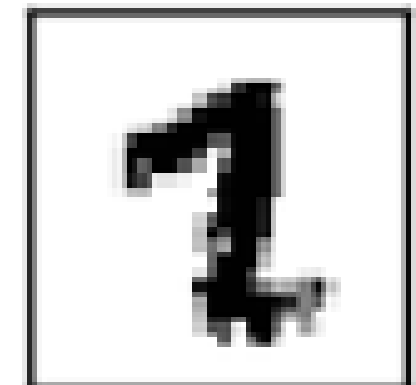
1: The University of Tokyo 2: RIKEN AIP



Introduction



Q: True or False!?



Q: 1,2 or 7 !?

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

Source: MNIST dataset
Lecun (1998)

Saying “I don’t know” can **prevent misclassification**.

Most theoretical works in this problem focused on binary case.

Only **Ramaswamy+ (2018)** considered confidence-based approach in multiclass case.

Contributions:

- Calibration condition to guarantee a surrogate loss in the **classifier-rejector approach**, which suggests the difficulty 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: $\{(x_i, y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} p(\mathbf{x}, y)$ $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$
Rejection cost: $c \in (0, 0.5)$ $y \in \mathcal{Y} = \{1, \dots, K\}$
 $g_i(\mathbf{x}): \mathcal{X} \rightarrow \mathbb{R}$
Find: Classifier: $f(\mathbf{x}) = \arg\max_{y \in \mathcal{Y}} g_y(\mathbf{x})$
Rejector: $r: \mathcal{X} \rightarrow \mathbb{R}$

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

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

$\mathcal{L}_{0-1-c}(r, f; \mathbf{x}, y)$ 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:

$$f^*(\mathbf{x}) = \arg \max_{y \in \mathcal{Y}} \eta_y(\mathbf{x}) \quad \eta_y(\mathbf{x}) = p(y|\mathbf{x}) \quad \text{Chow (1970)}$$
$$r^*(\mathbf{x}) = \max_{y \in \mathcal{Y}} \eta_y(\mathbf{x}) - (1 - c)$$

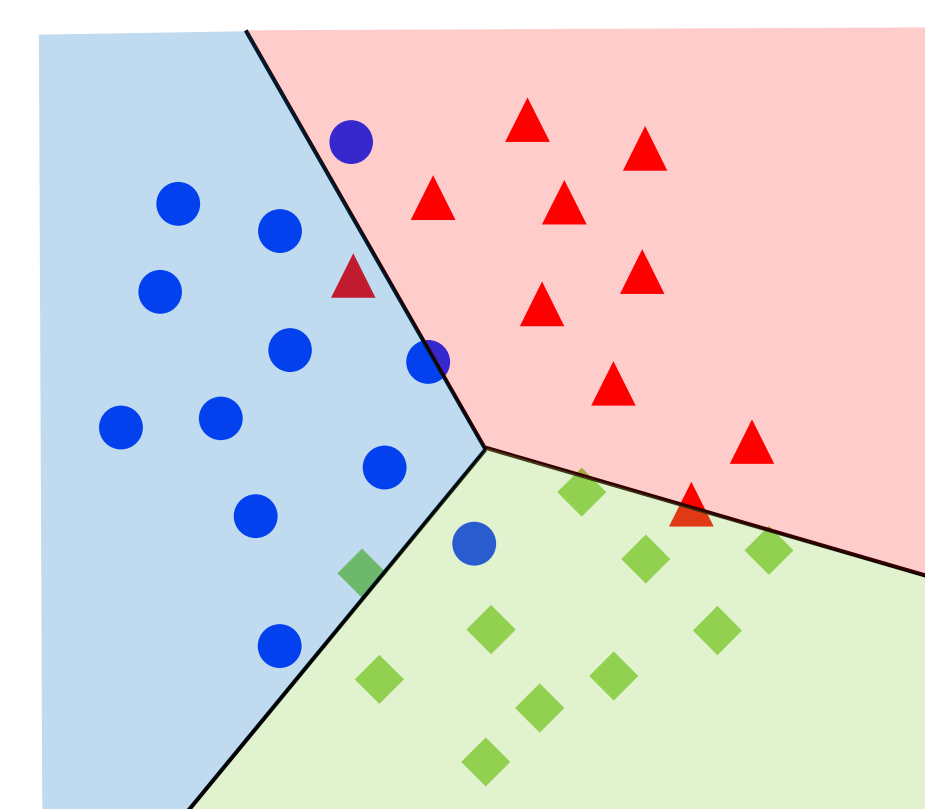
- (r, f) is **calibrated** if $R_{0-1-c}(r, f) = R_{0-1-c}(r^*, f^*)$
 - f is **classification-calibrated** if $f(\mathbf{x}) = f^*(\mathbf{x})$
 - r is **rejection-calibrated** if $\text{sign}[r(\mathbf{x})] = \text{sign}[r^*(\mathbf{x})]$
- If (r, f) is calibrated, r must be rejection-calibrated.

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

Classifier-rejector approach

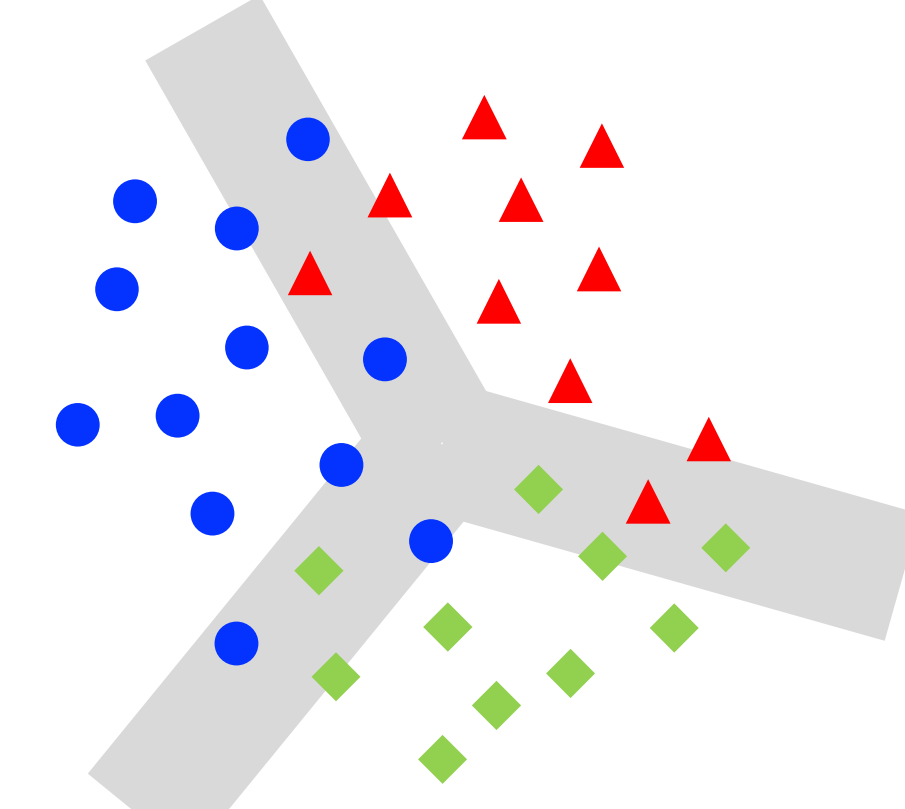
Cortes+ (2015, 2016)

Classifier and rejector are trained simultaneously



Classifier

(colored area indicates classifier prediction)



Rejector

(gray area indicates rejection area)

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

Corollary 5: (Necessary condition for rejection calibration)

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

$$\text{Condition (1)} \quad \sup_{\eta: \max_y \eta_y = 1-c} \left. \frac{\partial W(r, f_{\eta}^{\dagger}; \eta)}{\partial r} \right|_{r=0} = 0$$
$$\text{Condition (2)} \quad \inf_{\eta: \max_y \eta_y = 1-c} \left. \frac{\partial W(r, f_{\eta}^{\dagger}; \eta)}{\partial r} \right|_{r=0} = 0$$

Condition for false reject rate to be zero Condition for false accept rate to be zero

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

$$\sup_{\eta: \max_y \eta_y = 1-c} \left. \frac{\partial W(r, f_{\eta}^{\dagger}; \eta)}{\partial r} \right|_{r=0} = \inf_{\eta: \max_y \eta_y = 1-c} \left. \frac{\partial W(r, f_{\eta}^{\dagger}; \eta)}{\partial r} \right|_{r=0} = 0$$

Supremum and infimum values coincide under the same constraint.

When $\max_y \eta_y = 1 - c$

- Binary case: η can only be either $[1 - c, c]^{\top}$ or $[c, 1 - c]^{\top}$
- Multiclass case: η can be arbitrary. **Both conditions can be very different and do not hold simultaneously!**

Case study:

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

- Multiplicative pairwise comparison (MPC) loss:**

$$\mathcal{L}_{\text{MPC}}(r, f; \mathbf{x}, y) = \sum_{y' \neq y} \phi(\alpha(g_y(\mathbf{x}) - g_{y'}(\mathbf{x}))) \psi(-\alpha r(\mathbf{x})) + c \psi(\beta r(\mathbf{x}))$$

- Additive pairwise comparison (APC) loss:**

$$\mathcal{L}_{\text{APC}}(r, f; \mathbf{x}, y) = \sum_{y' \neq y} \phi(\alpha(g_y(\mathbf{x}) - g_{y'}(\mathbf{x}) - r(\mathbf{x}))) + c \psi(\beta r(\mathbf{x}))$$

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

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

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

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

In multiclass case, (α, β) that satisfies both conditions simultaneously **does not exist**.

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

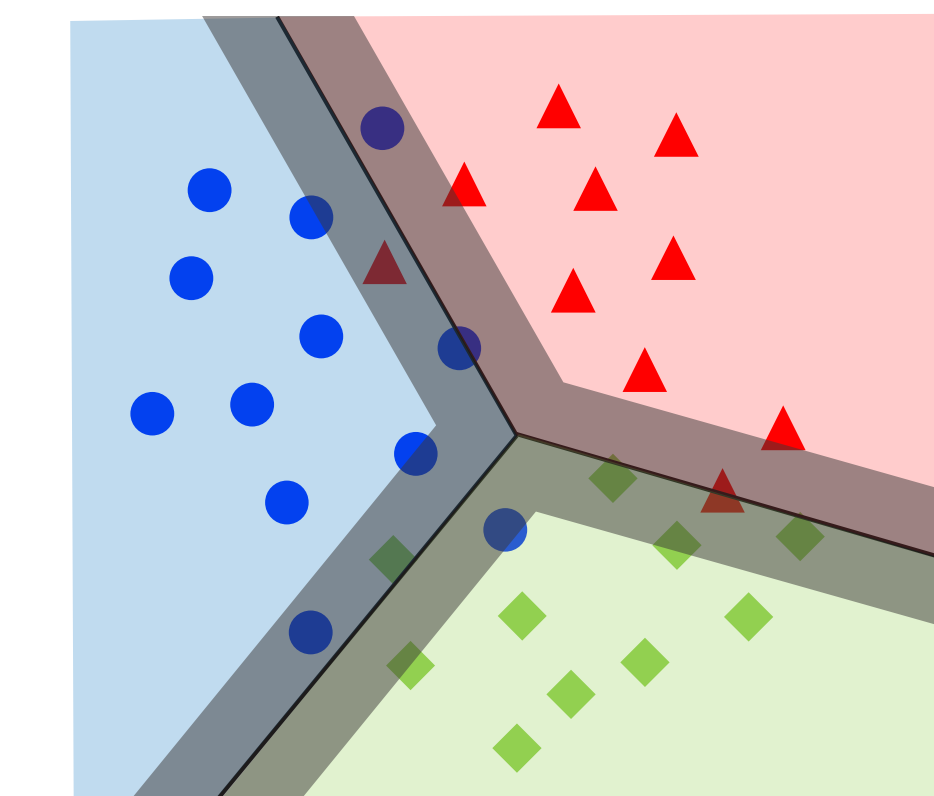
References

- [1] C. K. Chow. On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory, 1970.
- [2] C. Cortes, G. DeSalvo, and M. Mohri. Learning with rejection. ALT, 2015.
- [3] C. Cortes G. DeSalvo, and M. Mohri. Boosting with abstention. NeurIPS, 2016.
- [4] H.G. Ramaswamy, A. Tewari, and S. Agarwal. Consistent algorithms for multiclass classification with an abstain option. Electronic Journal of Statistics, 2018.
- [5] M. Yuan, M.H. Wegkamp. Classification methods with reject option based on convex risk minimization. JMLR, 2010.
- [6] 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:**
 $\mathcal{L}_{\text{CE}}(f; \mathbf{x}, y) = -g_y(\mathbf{x}) + \log \sum_{y' \in \mathcal{Y}} \exp(g_{y'}(\mathbf{x}))$
- One-versus-all (OVA) loss:**
 $\mathcal{L}_{\text{OVA}}(f; \mathbf{x}, y) = \phi(g_y(\mathbf{x})) + \sum_{y' \neq y} \phi(-g_{y'}(\mathbf{x}))$
- Rejector:**
 $r_f(\mathbf{x}) = \max_{y \in \mathcal{Y}} \Psi^{-1}(g(\mathbf{x})) - (1 - c)$

$$g(\mathbf{x}) = [g_1(\mathbf{x}), \dots, g_K(\mathbf{x})]^{\top}$$
$$\Psi^{-1}: \mathbb{R}^K \rightarrow [0, 1]^K \text{ Inverse link function}$$

$$\Psi_{y, \text{OVA}}^{-1}(g) = \frac{\phi'(-g_y)}{\phi'(-g_y) + \phi'(g_y)} \quad \Psi_{y, \text{CE}}^{-1}(g) = \frac{\exp(g_y)}{\sum_{y' \in \mathcal{Y}} \exp(g_{y'})}$$

See our paper for conditions of ϕ . Softmax function

Excess risk:

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

If ΔR_{0-1-c} can be upper-bounded by ΔR_{ℓ} ,

-> surrogate loss minimizer also minimizes ΔR_{0-1-c}

Excess risk bound of OVA loss:

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

Excess risk bound of CE loss:

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

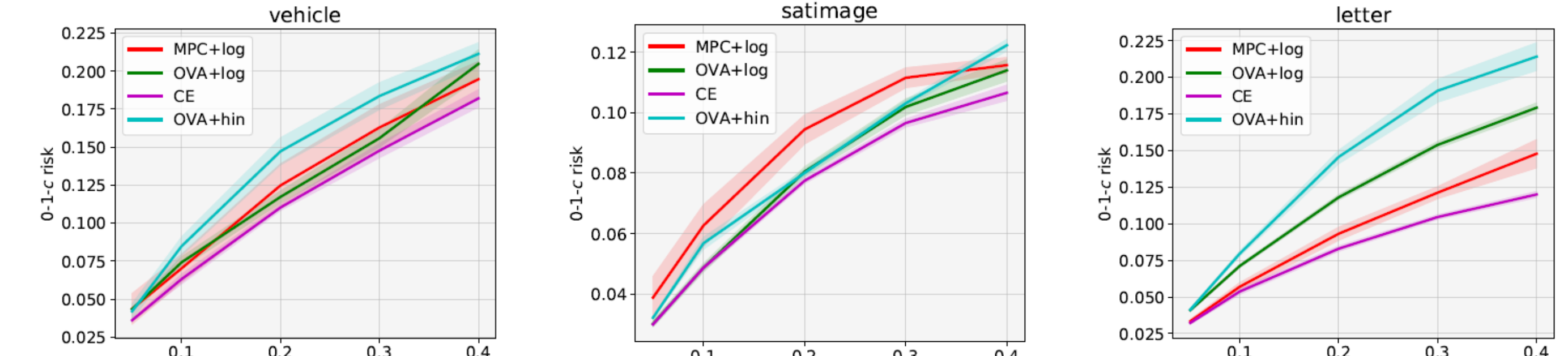
See our paper for more results, e.g., estimation error bound using Rademacher complexity.

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)