

On the Calibration of Multiclass Classification with Rejection

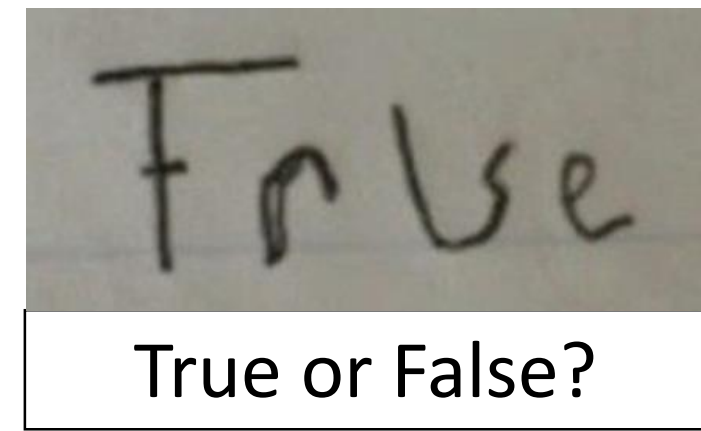
Chenri Ni¹ Nontawat Charoenphakdee^{1,2} Junya Honda^{1,2} Masashi Sugiyama^{2,1}



1: The University of Tokyo 2: RIKEN AIP



Introduction: Learning with rejection



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

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) only focused on specific types of non-differentiable losses.

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

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

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

Find: Classifier: $f(x) = \arg\max_{y \in \mathcal{Y}} g_y(x) \in \mathcal{Y}$

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

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

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

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

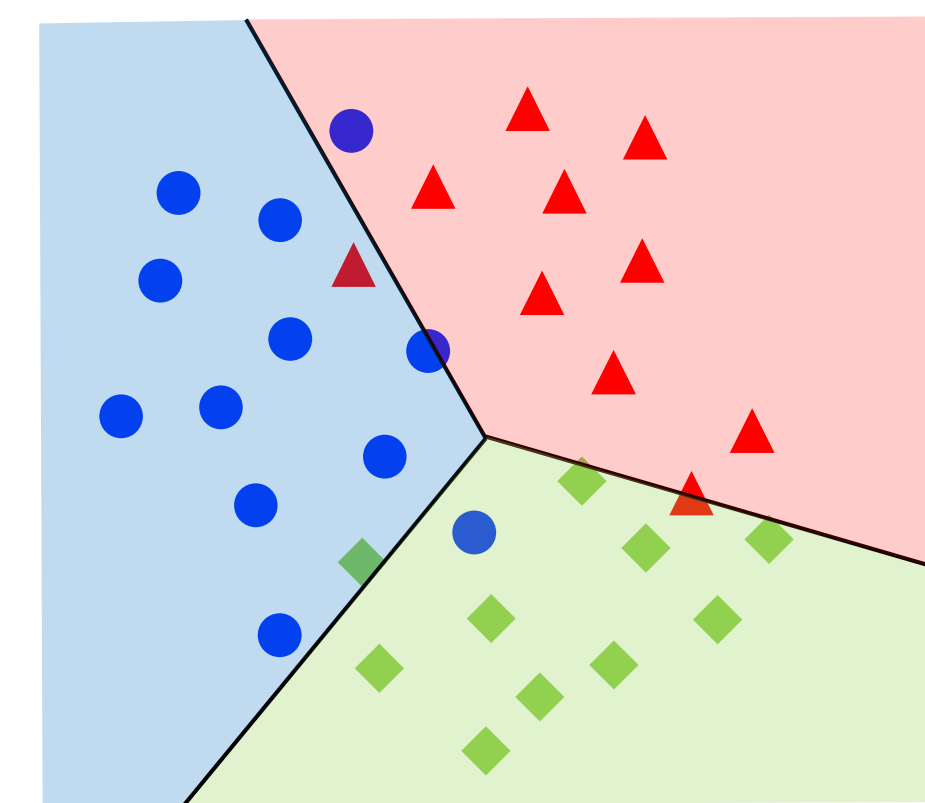
$$f^*(x) = \arg \max_{y \in \mathcal{Y}} \eta_y(x) \quad \eta_y(x) = p(y|x) \quad \text{Chow (1970)}$$
$$r^*(x) = \max_{y \in \mathcal{Y}} \eta_y(x) - (1 - c)$$

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

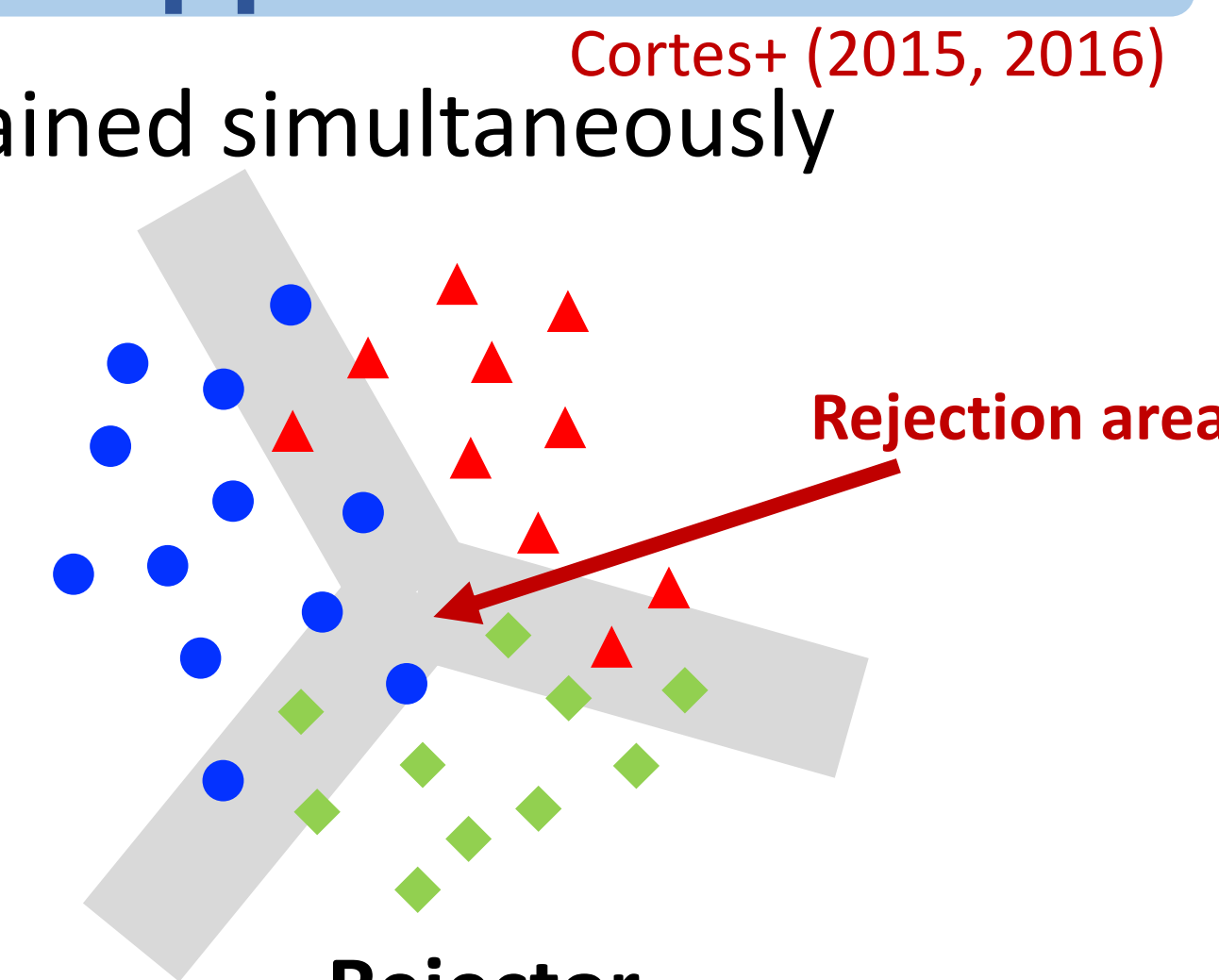
A minimizer of a surrogate loss should give a calibrated (r, f) .

Classifier-rejector approach

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.

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

Corollary 5: (Necessary condition for rejection calibration)

For $\mathcal{L}(r, f; x, y)$ 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:

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

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

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

Case study:

• **Multiplicative pairwise comparison (MPC) loss:**

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

• **Additive pairwise comparison (APC) loss:**

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

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 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 $\phi(z) = \psi(z) = \log(1 + \exp(-z))$.

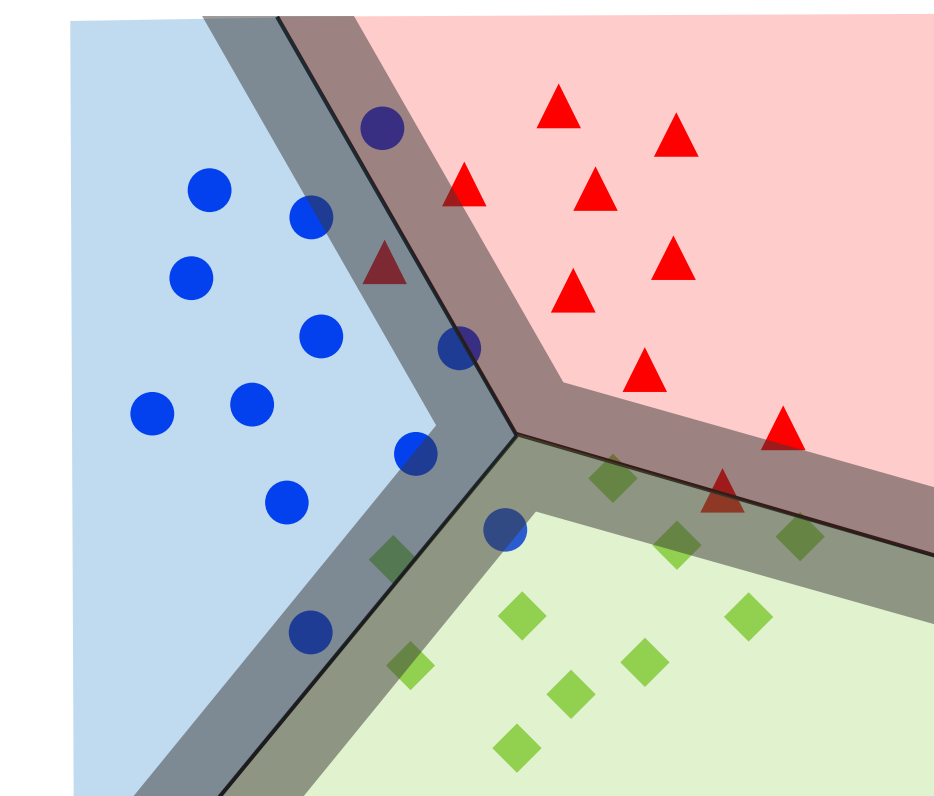
References

- [1] K. 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:**
 $\mathcal{L}_{\text{CE}}(f; x, y) = -g_y(x) + \log \sum_{y' \in \mathcal{Y}} \exp(g_{y'}(x))$
- One-versus-all (OVA) loss:**
 $\mathcal{L}_{\text{OVA}}(f; x, y) = \phi(g_y(x)) + \sum_{y' \neq y} \phi(-g_{y'}(x))$
- Rejector:**
 $r_f(x) = \max_{y \in \mathcal{Y}} \Psi^{-1}(g(x)) - (1 - c)$
 $g(x) = [g_1(x), \dots, g_K(x)]^\top$
 $\Psi^{-1}: \mathbb{R}^K \rightarrow [0, 1]^K$ 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 on ϕ . Softmax function

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

Excess risk:

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

Excess risk bound of OVA loss:

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

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

Excess risk bound of CE loss:

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

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.

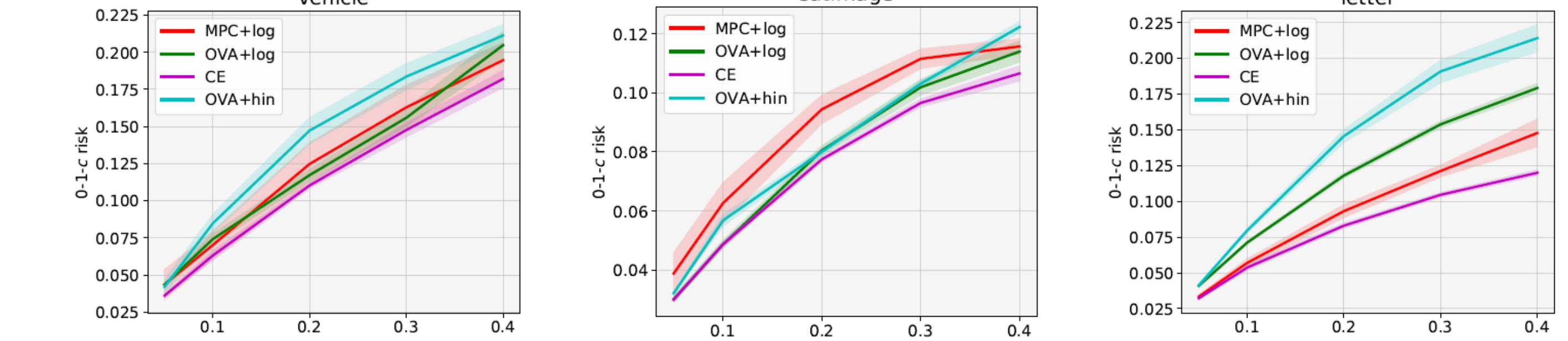
See our paper for estimation error bound using Rademacher complexity.

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)