A General Online Algorithm for Optimizing Complex Performance Metrics

Wojciech Kotłowski[1] Marek Wydmuch[1] Erik Schultheis[2] Rohit Babbar[2,3] Krzysztof Dembczyński[1,4]

[1]Poznan University of Technology, Poland [2]Aalto University, Helsinki, Finland [3]University of Bath, UK [4]Yahoo Research, New York, USA

1. Online classification

- Classification: $\boldsymbol{x} \in \mathcal{X} \to \boldsymbol{y} \in \mathcal{Y}$, $(\boldsymbol{x}, \boldsymbol{y}) \sim \mathbb{P}$.
- $\mathbb{P}_{\mathcal{X}}$ – marginal distribution of $\boldsymbol{x}$
- $\boldsymbol{\eta}(\boldsymbol{x})$ – label conditional probability, $\boldsymbol{\eta}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}[\boldsymbol{y}]$.
- **Multi-class classification**:
  - $\boldsymbol{y} \in \{\boldsymbol{y}' \in \{0,1\}^m \colon \|\boldsymbol{y}'\|_1 = 1\}$ – a vector one-hot encoding of one out of m classes,
  - $\boldsymbol{\eta}(\boldsymbol{x}) = (\eta_1(\boldsymbol{x}), \ldots, \eta_m(\boldsymbol{x}))$ – label conditional distribution with $\eta_j(\boldsymbol{x}) = \mathbb{P}(y_j = 1|\boldsymbol{x})$
- **Multi-label classification**:
  - $\boldsymbol{y} := \{0,1\}^m$ – a vector of relevant labels
  - $\boldsymbol{\eta}(\boldsymbol{x}) = (\eta_1(\boldsymbol{x}), \ldots, \eta_m(\boldsymbol{x}))$ – a vector of **label marginal conditional probabilities**, $\eta_j(\boldsymbol{x}) = \mathbb{P}(y_j = 1|\boldsymbol{x})$.
- **Online protocol**:

Input: a sequence $(\boldsymbol{x}^n, \boldsymbol{y}^n) = ((\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n))$
for $t = 1, \ldots, n$ **do**
  Observe input instance $\boldsymbol{x}_t$ drawn from $\mathbb{P}_{\mathcal{X}}$
  Receive conditional probability estimate $\widehat{\boldsymbol{\eta}}_t(\boldsymbol{x}_t)$
  Predict label $\widehat{\boldsymbol{y}}_t$ based on $\widehat{\boldsymbol{\eta}}_t(\boldsymbol{x}_t)$
  Receive true label $\boldsymbol{y}_t$ drawn from $\mathbb{P}(\cdot|\boldsymbol{x}_t)$
Evaluate based on $\psi(\boldsymbol{y}^n, \widehat{\boldsymbol{y}}^n)$

- **Goal**: Maximize performance metric $\psi(\boldsymbol{y}^n, \widehat{\boldsymbol{y}}^n)$

2. Confusion matrices

Binary $\boldsymbol{C}^{2\times 2}$:

y \ $\hat{y}$	0	1
0	TN	FP
1	FN	TP

Multi-class $\boldsymbol{C}^{m\times m}$:

y \ $\hat{y}$	1	2	…	m
1	TP_1	$\mathrm{E}_{1,2}$	…	$\mathrm{E}_{1,m}$
2	$\mathrm{E}_{2,1}$	TP_2	…	$\mathrm{E}_{2,m}$
⋮	⋮	⋮	⋱	⋮
m	$\mathrm{E}_{m,1}$	$\mathrm{E}_{m,2}$	…	TP_m

Multi-label $\boldsymbol{C}^{m\times 2\times 2}$:

$$\left[\begin{array}{cc|cc} & \hat{y}_1 & & \hat{y}_2 \\ y_1 & 0 \quad 1 & y_2 & 0 \quad 1 \\ 0 & \mathrm{TN}_1 \quad \mathrm{FP}_1 & 0 & \mathrm{TN}_2 \quad \mathrm{FP}_2 \\ 1 & \mathrm{FN}_1 \quad \mathrm{TP}_1 & 1 & \mathrm{FN}_2 \quad \mathrm{TP}_2 \end{array}, \ldots, \begin{array}{cc} & \hat{y}_m \\ y_m & 0 \quad 1 \\ 0 & \mathrm{TN}_m \quad \mathrm{FP}_m \\ 1 & \mathrm{FN}_m \quad \mathrm{TP}_m \end{array}\right]$$

3. Complex performance metrics

We focus on the online maximization of performance metrics that do **not decompose into a sum over instances** but are general **functions of the confusion matrix**.

Examples of binary and multi-class confusion matrix measures.

Metric	$\psi(\boldsymbol{C}^{2\times 2})$	$\psi(\boldsymbol{C}^{m\times m})$
Balanced Acc.	$\frac{\mathrm{tp}}{2(\mathrm{tp}+\mathrm{fn})} + \frac{\mathrm{tn}}{2(\mathrm{tn}+\mathrm{fp})}$	$\sum_{i=1}^m \frac{C_{ii}}{m\sum_{j=1}^m C_{ij}}$
Recall	$\frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fn}}$	micro- or macro-average
Precision	$\frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fp}}$	micro- or macro-average
F_β-measure	$\frac{(1+\beta^2)\mathrm{tp}}{(1+\beta^2)\mathrm{tp}+\beta^2\mathrm{fn}+\mathrm{fp}}$	micro- or macro-average
G-mean	$\sqrt{\frac{\mathrm{tp}\cdot\mathrm{tn}}{(\mathrm{tp}+\mathrm{fn})(\mathrm{tn}+\mathrm{fp})}}$	$\left(\prod_{j=1}^m \frac{C_{jj}}{\sum_{i=1}^m C_{ji}}\right)^{1/m}$
H-mean	$2\left(\frac{\mathrm{tp}+\mathrm{fn}}{\mathrm{tp}} + \frac{\mathrm{tn}+\mathrm{fp}}{\mathrm{tn}}\right)^{-1}$	$m\left(\sum_{j=1}^m \frac{\sum_{i=1}^m C_{ji}}{C_{jj}}\right)^{-1}$
Q-mean	$1-\sqrt{\frac{1}{2}\left(\left(\frac{\mathrm{fn}}{\mathrm{tp}+\mathrm{fn}}\right)^2+\left(\frac{\mathrm{fp}}{\mathrm{tn}+\mathrm{fp}}\right)^2\right)}$	$1-\sqrt{\frac{1}{m}\sum_{j=1}^m\left(1-\frac{C_{jj}}{\sum_{i=1}^m C_{ji}}\right)^2}$

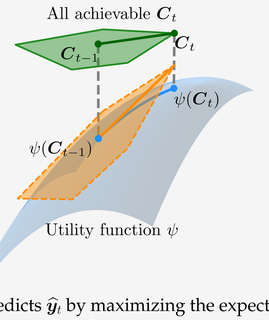
Micro- and macro-averages of metrics for $\boldsymbol{C}^{m\times 2\times 2}$:

$$\text{Micro-}\psi(\boldsymbol{C}) := \psi\left(\frac{1}{m}\sum_{j=1}^m \boldsymbol{C}_j\right), \quad \text{Macro-}\psi(\boldsymbol{C}) := \frac{1}{m}\sum_{j=1}^m \psi(\boldsymbol{C}_j)\ .$$

4. Online Metric Maximi. Algo. (OMMA)

Under assumption that the utility $\psi(\boldsymbol{C})$ is **differentiable** in $\boldsymbol{C}$:

1: **Initialization:** confusion matrix $\boldsymbol{C}_0$
2: **for** $t = 1, \ldots, n$ **do**
3: Receive input $\boldsymbol{x}_t$ and probability estimate $\widehat{\boldsymbol{\eta}}_t(\boldsymbol{x}_t)$
4: Predict $\widehat{\boldsymbol{y}}_t = \operatorname{argmax}_{\widehat{\boldsymbol{y}}} \nabla\psi(\boldsymbol{C}_{t-1}) \cdot \left(\mathbb{E}_{\boldsymbol{y}_t\sim\widehat{\boldsymbol{\eta}}_t(\boldsymbol{x}_t)}[\boldsymbol{C}(\boldsymbol{y}_t, \widehat{\boldsymbol{y}})]\right)$
5: Receive label $\boldsymbol{y}_t$ and update $\boldsymbol{C}_t = \frac{t-1}{t}\boldsymbol{C}_{t-1} + \frac{1}{t}\boldsymbol{C}(\boldsymbol{y}_t, \widehat{\boldsymbol{y}}_t)$



At trial t, OMMA predicts $\widehat{\boldsymbol{y}}_t$ by maximizing the expected **linearized** utility at $\boldsymbol{C}_{t-1}$ based on probability estimates .

5. Theoretical results

If $\psi(\boldsymbol{C})$ is **concave, Lipschitz, and smooth**:

- With $\boldsymbol{\eta}(\boldsymbol{x})$ available, the regret of OMMA is of order $\mathcal{O}(\frac{\ln n}{n})$.
- Otherwise, the regret is bounded by $\mathbb{E}\left[\|\boldsymbol{\eta}(\boldsymbol{x}) - \widehat{\boldsymbol{\eta}}\boldsymbol{x})\|\right]$.
- If the estimation error converges to zero with $n \to \infty$, OMMA becomes a **no-regret** learning algorithm.

For some ψ, Lipschitzness and smoothness might **not hold globally**:

- These properties are invoked only along the parameter path of the algorithm, $\{\boldsymbol{C}_t\}_{t=1}^n$.
- Many ψ are Lipschitz and smooth when label frequencies are bounded away from zero
- $\implies$ adding a small value λ to $\boldsymbol{C}$ stabilizes the algorithm.

$\implies$ OMMA is a **robust** approach for online learning scenarios.

6. Example: binary classification

- $y_t \in \{0,1\}$, and $\eta(\boldsymbol{x}_t) = P(y_t = 1|\boldsymbol{x}_t)$.
- Because:

$$\mathbb{E}_{y_t\sim\widehat{\eta}(\boldsymbol{x}_t)}[\boldsymbol{C}(y_t, \widehat{y})] = \begin{bmatrix} (1-\widehat{\eta}(\boldsymbol{x}_t))(1-\widehat{y}) & (1-\widehat{\eta}(\boldsymbol{x}_t))\widehat{y} \\ \widehat{\eta}(\boldsymbol{x}_t)(1-\widehat{y}) & \widehat{\eta}(\boldsymbol{x}_t)\widehat{y} \end{bmatrix},$$

line 3 of the algorithm boils down to maximizing a cost-sensitive classification accuracy $\widehat{y}(\alpha\widehat{\eta}(\boldsymbol{x}_t) - \beta)$, with:

$$\alpha = \nabla_{11} + \nabla_{00} - \nabla_{01} - \nabla_{10}, \quad \beta = \nabla_{00} - \nabla_{01}\ .$$

- All practical utilities are non-decreasing with TP and TN ($\nabla_{11}, \nabla_{00} \geq 0$), and non-increasing with FN and FP ($\nabla_{01}, \nabla_{10} \leq 0$).
- $a \geq 0 \to$ maximizing $\widehat{y}(\alpha\widehat{\eta}(\boldsymbol{x}_t) - \beta)$ boils down to choosing $\widehat{y}_t = 1$ when $\widehat{\eta}(\boldsymbol{x}_t) > \beta/\alpha$ – well-known rule for maximizing the population-level version of ψ (OMMA **mimics** this optimal rule).

Read the paper

Try our library

https://github.com/mwydmuch/xCOLUMNs

7. Experimental results

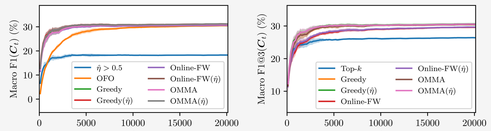
Results of different online algorithms on multi-class and multi-label problems compared to OMMA (all use the same $\widehat{\boldsymbol{\eta}}(\boldsymbol{x})$):

	LEDGAR-LEXGLUE ($m = 100, n = 10000$)							FLICKR ($m = 195, n = 24154$)						
	Macro			Multi-class			Micro	Macro					
Method	F1	F1@3	Rec.@3	Pr.@3	G-mean	H-mean	Q-mean	F1	F1	F1@3	Rec.@3	Pr.@3	G-mean	H-mean
Top-k / $\eta_i > 0.5$	79.06	51.80	92.04	38.88	0.00	0.00	69.21	29.46	18.27	26.39	38.96	21.38	27.08	20.02
OFO	×	×	×	×	×	×	×	**41.05**	30.46	×	×	×	×	×
Greedy	79.30	78.08	93.26	91.17	×	×	×	×	30.90	30.42	**46.41**	57.55	83.39	**83.37**
Online-FW*	79.22	70.94	93.30	54.52	62.31	75.81	**74.59**	**41.05**	30.60	29.58	46.38	58.66	83.37	*83.28*
Online-FW(η_i)*	79.22	73.85	*93.38*	49.63	*78.02*	**77.58**	74.50	*41.02*	**31.17**	29.65	46.28	25.73	83.37	83.20
OMMA (ours)	79.28	*78.10*	93.26	**91.41**	77.48	74.62	**74.59**	41.01	30.90	30.39	**46.41**	**58.35**	*83.39*	**83.37**
OMMA($\widehat{\eta}_i$) (ours)	**79.34**	**78.22**	**93.39**	90.10	**78.03**	*76.08*	*74.53*	*41.02*	*31.15*	**30.55**	46.33	55.56	**83.41**	83.23

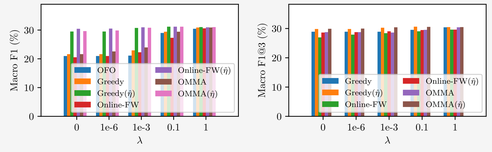
	RCV1X ($m = 2456, n = 155962$)							AMAZONCAT ($m = 13330, n = 306784$)						
	Micro	Macro						Micro	Macro					
Method	F1	F1	F1@3	Rec.@3	Pr.@3	G-mean	H-mean	F1	F1	F1@3	Rec.@3	Pr.@3	G-mean	H-mean
Top-k / $\eta_i > 0.5$	68.57	11.29	5.34	4.59	13.24	16.01	*12.36*	*67.77*	28.76	14.98	11.18	30.98	*33.93*	*30.03*
OFO	**69.83**	20.26	×	×	×	×	×	*70.38*	39.60	×	×	×	×	×
Greedy	×	**20.80**	*16.01*	21.20	**30.91**	**69.07**	**67.04**	×	44.20	43.64	57.81	**54.00**	**80.86**	**80.04**
Online-FW*	**69.83**	19.82	15.33	21.09	19.88	**69.07**	**67.04**	**70.61**	42.42	40.20	57.74	40.10	**80.86**	**80.04**
Online-FW(η_i)*	*69.79*	20.40	15.55	*22.21*	22.17	*69.06*	**67.04**	70.34	*47.32*	*44.63*	*58.87*	49.68	**80.86**	**80.04**
OMMA (ours)	69.77	20.57	15.87	21.08	30.39	**69.07**	**67.04**	69.90	43.04	42.14	57.80	*52.04*	**80.86**	**80.04**
OMMA($\widehat{\eta}_i$) (ours)	69.71	*20.71*	**16.07**	**22.23**	30.38	*69.06*	**67.04**	70.02	**47.67**	**45.06**	**58.89**	51.52	**80.86**	**80.04**

× – metric is not supported by an algorithm, * – needs to store all previously seen instances.

Comparison of the incremental performance of the online algorithms on the Flickr dataset:



Impact of metric smoothing parameter λ on the results of the online algorithms on the Flickr dataset:



The 41st International Conference on Machine Learning (ICML), July 21–27, 2024, Vienna, Austria