Supplemental Material

Paper ID: 518

A Appendix

A.1 Convex Relaxation and Dual of Problem (11)

Since problem (11) is a mixed integer problem regarding $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$, it is hard to directly optimize. Motivated by (Tan *et al.* 2010), we apply convex relaxation and Lagrange dual to make some transformations.

Firstly, we introduce dual variables $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^{N_h}$ for the hingle loss constraint:

$$\ell(\boldsymbol{\mu}, \boldsymbol{\eta}; (\mathbf{x}_i, y_i)) = \max(0, 1 - y_i \boldsymbol{\mu}^T (\boldsymbol{\eta} \odot \mathbf{x}_i)). \quad (21)$$

As to problem (11), we can get the Lagrangian function of the inner problem w.r.t $\boldsymbol{\mu}$:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\mu}, \ell, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu_{h-1}})^T \boldsymbol{\Sigma}_h(\boldsymbol{\eta})^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu_{h-1}}) \\ &+ \frac{C}{q}\sum_{i=1}^{N_h} D_i \ell(\boldsymbol{\mu}, \boldsymbol{\eta}; (\mathbf{x}_i, y_i))^q + \sum_{i=1}^{N_h}(-\boldsymbol{\beta}_i \ell(\boldsymbol{\mu}, \boldsymbol{\eta}; (\mathbf{x}_i, y_i))) \\ &+ \sum_{i=1}^{N_h} \alpha_i (1 - y_i(\boldsymbol{\mu} \odot \boldsymbol{\eta})^T \mathbf{x}_i - \ell(\boldsymbol{\mu}, \boldsymbol{\eta}; (\mathbf{x}_i, y_i))) \end{aligned} \quad (22)$$

By taking derivative over variables $\boldsymbol{\mu}$ and ℓ, we get:

$$\begin{aligned} &\nabla_{\boldsymbol{\mu}} \mathcal{L} = \boldsymbol{\Sigma}(\boldsymbol{\eta})^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_{h-1}) - \sum_{i=1}^{N_h} \alpha_i y_i (\mathbf{x}_\mathbf{i} \odot \boldsymbol{\eta}) = 0, \\ &\nabla_{\ell_i} \mathcal{L} = CD_i - \alpha_i - \beta_i = 0, \alpha_i, \beta_i \geq 0, \text{ for q=1 }, \\ &\nabla_{\ell_i} \mathcal{L} = CD_i \ell_i - \alpha_i - \beta_i = 0, \beta_i = 0, \text{ for q=2 }. \end{aligned}$$

With some transformations, then:

$$\begin{aligned} &\boldsymbol{\mu} - \boldsymbol{\mu}_{h-1} = \boldsymbol{\Sigma}(\boldsymbol{\eta}) \sum_{i=1}^{N_h} \alpha_i y_i (\mathbf{x}_i \odot \boldsymbol{\eta}), \\ &0 \leq \alpha_i \leq CD_i, \text{ for q=1 }, \\ &\ell_i = \alpha_i/(CD_i), \text{ for q=2 }. \end{aligned}$$

Let $\mathbf{g}(\boldsymbol{\alpha}, \boldsymbol{\eta}) := \sum_{i=1}^{N_h} \alpha_i y_i (\mathbf{x}_i \odot \boldsymbol{\eta})$, $\mathcal{A} := \{\boldsymbol{\alpha} \in \mathbb{R}^{N_h} | 0 \leq \alpha_i \leq U, \forall i \in [N_h]\}$ is the domain of $\boldsymbol{\alpha}$ (here, $U = CD_i$ for $q = 1$ and $U = \infty$ for $q = 2$), then we can get the dual of inner problem (11) as:

$$\begin{aligned} \max_{\boldsymbol{\alpha} \in \boldsymbol{\Lambda}} &- \frac{1}{2}\mathbf{g}(\boldsymbol{\alpha}, \boldsymbol{\eta})^T \boldsymbol{\Sigma}(\boldsymbol{\eta}) \mathbf{g}(\boldsymbol{\alpha}, \boldsymbol{\eta}) - \frac{q-1}{2C}\sum_{i=1}^{N_h} \frac{\alpha_i^2}{D_i} \\ &+ \sum_{i=1}^{N_h} \alpha_i - \boldsymbol{\mu}_{h-1}^T \mathbf{g}(\boldsymbol{\alpha}, \boldsymbol{\eta}), \end{aligned} \quad (23)$$

We define objective of (23) as $f(\boldsymbol{\alpha}, \boldsymbol{\eta})$ for convenience. Problem (11) can be reformulated as a minmax problem:

$$\min_{\boldsymbol{\eta} \in \boldsymbol{\Lambda}} \max_{\boldsymbol{\alpha} \in \mathcal{A}} f(\boldsymbol{\alpha}, \boldsymbol{\eta}), \quad (24)$$

This problem is also a mixed integer problem, but we have the following property according to minmax inequality (Sion 1958):

$$\min_{\boldsymbol{\eta} \in \boldsymbol{\Lambda}} \max_{\boldsymbol{\alpha} \in \mathcal{A}} f(\boldsymbol{\alpha}, \boldsymbol{\eta}) \geq \max_{\boldsymbol{\alpha} \in \mathcal{A}} \min_{\boldsymbol{\eta} \in \boldsymbol{\Lambda}} f(\boldsymbol{\alpha}, \boldsymbol{\eta}), \quad (25)$$

The latter problem of (25) provides a lower bound to problem (24) and it is also a convex problem. By introducing a variable θ, we can transform the problem into :

$$\max_{\theta \in \mathbb{R}, \boldsymbol{\alpha} \in \mathcal{A}} \theta, \text{ s.t. } \theta \leq f(\boldsymbol{\alpha}, \boldsymbol{\eta}), \forall \eta \in \boldsymbol{\Lambda}. \quad (26)$$

A.2 Solving for the Primal of Problem (15)

We prove that problem (16) is the primal of problem (15).

Let $\Omega(\mathbf{w}) = \frac{1}{2}(\sum_{k=1}^{K} \|\mathbf{w}_k - \mathbf{w}_k^{h-1}\|)^2$. Define second order crone $\mathcal{Q}_r = \{(\mathbf{u}, \mathbf{v}) \in \mathbb{R}^{r+1}, \|\mathbf{u}\|_2 \leq v\}$. Let $\mathbf{z}_k = \|\mathbf{w}_k - \mathbf{w}_k^{h-1}\|$, then $\Omega(\mathbf{w}) = \frac{1}{2}\mathbf{z}^2$, where $\mathbf{z} = \sum_{k=1}^{K} \mathbf{z}_k, \mathbf{z}_k \geq \mathbf{0}$ and $\mathbf{z} \geq \mathbf{0}$. Then problem (16) can be reformulated as:

$$\begin{aligned} \min_{\mathbf{z}, \mathbf{w}_k, \ell} &\frac{1}{2}\mathbf{z}^2 + \frac{C}{q}\sum_{i=1}^{N_h} D_i \ell(\mathbf{w}_k, \boldsymbol{\eta})^q, \\ \text{s.t. } &\sum_{k=1}^{K} \mathbf{z}_k \leq \mathbf{z}, (\mathbf{w}_k - \mathbf{w}_k^{h-1}, \mathbf{z}_k) \in \mathcal{Q}_r, \end{aligned} \quad (27)$$

We introduce $\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \boldsymbol{\delta}, \boldsymbol{\epsilon}$ here. With $\boldsymbol{\delta}, \boldsymbol{\epsilon}$ and the constraints on second order crone $\mathcal{Q}_r$, we point out that $\boldsymbol{\delta}_k^T(\mathbf{w}_k -$

$\mathbf{w}_k^{h-1}) + \epsilon_k \mathbf{z}_k$ along with $\|\boldsymbol{\delta}_k\| \leq \epsilon_k$ equals original constraints on $\mathcal{Q}_r$ with Lagrangian multiplier. Now Lagrangian function can be written as:

$$\mathcal{L}(\mathbf{z}, \mathbf{w}_k, \ell, \boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \boldsymbol{\delta}, \boldsymbol{\epsilon}) = \frac{1}{2}\mathbf{z}^2 + \frac{C}{q}\sum_{i=1}^{N_h} D_i \ell_i^q + \sum_{i=1}^{N_h}(-\beta_i \ell_i)$$
$$- \sum_{k=1}^{K}(\boldsymbol{\delta}_k^T(\mathbf{w}_k - \mathbf{w}_k^{h-1}) + \epsilon_k \mathbf{z}_k) + \gamma(\sum_{k=1}^{K}\mathbf{z}_k - \mathbf{z})$$
$$+ \sum_{i=1}^{N_h}\alpha_i(1 - y_i\sum_{k=1}^{K}\mathbf{w}_k^T\widehat{\mathbf{x}}_i^k - \ell_i). \quad (28)$$

Taking derivatives w.r.t $\mathbf{z}, \mathbf{w}_k, \ell_i$, the KKT condition is as follows:

$$\nabla_{\mathbf{z}}\mathcal{L} = \mathbf{z} - \gamma = 0,$$
$$\nabla_{\mathbf{z}_k}\mathcal{L} = \gamma - \epsilon_k = 0,$$
$$\nabla_{\mathbf{w}_k}\mathcal{L} = -\sum_{i=1}^{N_h}\alpha_i y_i \widehat{\mathbf{x}}_i^k - \boldsymbol{\delta}_k = 0,$$
$$\nabla_{\ell_i}\mathcal{L} = CD_i - \alpha_i - \beta_i = 0, \alpha_i, \beta_i \geq 0, \text{ for q=1 },$$
$$\nabla_{\ell_i}\mathcal{L} = CD_i\ell_i - \alpha_i - \beta_i = 0, \beta_i = 0, \text{ for q=2 },$$
$$\|\boldsymbol{\delta}_k\| \leq \epsilon_k.$$

Substituting all the equations back into Lagrangian function, we have

$$\mathcal{L}(\mathbf{z}, \mathbf{w}_k, \ell, \boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \boldsymbol{\delta}, \boldsymbol{\epsilon}) = -\frac{1}{2}\gamma^2 - \frac{q-1}{2C}\sum_{i=1}^{N_h}\frac{\alpha_i^2}{D_i} + \sum_{i=1}^{N_h}\alpha_i$$
$$+ \sum_{k=1}^{K}\boldsymbol{\delta}^k \mathbf{w}_k^{h-1}. \quad (29)$$

Let $\mathcal{A} := \{\boldsymbol{\alpha} \in \mathbb{R}^{N_h} | 0 \leq \alpha_i \leq U\}$ be the domain of $\boldsymbol{\alpha}$ where $U = CD_i$ for $q = 1$ and $U = \infty$ for $q = 2$. We then rewrite dual problem of Lagrangian:

$$\max_{\gamma\in\mathbb{R}, \boldsymbol{\alpha}\in\mathcal{A}} \mathcal{L}(\mathbf{z}, \mathbf{w}_k, \ell, \boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \boldsymbol{\delta}, \boldsymbol{\epsilon})$$
$$\text{s.t.} \|\sum_{k=1}^{K}\alpha_i y_i \widehat{\mathbf{x}}_i^k\| \leq \gamma, k = 1, \ldots, K \quad (30)$$

Let $\theta := \mathcal{L}(\mathbf{z}, \mathbf{w}_k, \ell, \boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \boldsymbol{\delta}, \boldsymbol{\epsilon})$ and $\mathbf{g}(\boldsymbol{\alpha}, \boldsymbol{\eta}_k) = \sum_{k=1}^{K}\alpha_i y_i(\widehat{\mathbf{x}}_i \odot \boldsymbol{\eta}_k)$. We further define $f(\boldsymbol{\alpha}, \boldsymbol{\eta}_k) = -\frac{1}{2}\mathbf{g}(\boldsymbol{\alpha}, \boldsymbol{\eta}_k)^T\mathbf{g}(\boldsymbol{\alpha}, \boldsymbol{\eta}_k) - \frac{q-1}{2C}\sum_{i=1}^{N_h}\frac{\alpha_i^2}{D_i} + \sum_{i=1}^{N_h}\alpha_i - (\mathbf{w}_k^{h-1})^T\mathbf{g}(\boldsymbol{\alpha}, \boldsymbol{\eta}_k)$. Then,

$$\max_{\theta, \boldsymbol{\alpha}\in\mathcal{A}} \theta, \text{ s.t. } \theta \leq f(\boldsymbol{\alpha}, \boldsymbol{\eta_k}), k = 1, \ldots, K. \quad (31)$$

Since $\widehat{\mathbf{x}}_i = \boldsymbol{\Sigma}_k^{\frac{1}{2}}\mathbf{x}_i$, with some transformation, we can get that (31) is equivalent to (15).

A.3 Conjugate Dual of Problem (17)

Problem (17) can be written as:

$$\min_{\mathbf{w}} \Omega(\mathbf{w}) + C\sum_{i=1}^{N_h} L_i(\mathbf{w}^T\widehat{\mathbf{x}}_i). \quad (32)$$

Let $p_i := \mathbf{w}^T\widehat{\mathbf{x}}_i$, (32) can be reformulated as:

$$\min_{\mathbf{w}} \Omega(\mathbf{w}) + C\sum_{i=1}^{N_h} L_i(p_i), \text{ s.t. } p_i = \mathbf{w}^T\widehat{\mathbf{x}}_i, i = 1, \ldots, N_h \quad (33)$$

Now the Lagrangian function:

$$\mathcal{L}(\mathbf{w}, \mathbf{p}, \boldsymbol{\alpha}) = \Omega(\mathbf{w}) + C\sum_{i=1}^{N_h} L_i(\mathbf{p}_i) + C\sum_{i=1}^{N_h}\alpha_i(p_i - \mathbf{w}^T\widehat{\mathbf{x}}_i)$$
$$= \Omega(\mathbf{w}) - C\sum_{i=1}^{N_h}\alpha_i\mathbf{w}^T\widehat{\mathbf{x}}_i + C\sum_{i=1}^{N_h}(L_i(p_i) + \alpha_i p_i). \quad (34)$$

Let $\mathbf{z}(\boldsymbol{\alpha}) = C\sum_{i=1}^{N_h}\alpha_i\widehat{\mathbf{x}}_i$. If we decouple $\mathbf{w}$ and $\mathbf{p}$, and minimize Lagrangian function w.r.t $\mathbf{w}$ and $\mathbf{p}$, then:

$$\min_{\mathbf{w},\mathbf{p}} \mathcal{L}(\mathbf{w}, \mathbf{p}, \boldsymbol{\alpha})$$
$$= \min_{\mathbf{w}}(\Omega(\mathbf{w}) - \mathbf{w}^T\mathbf{z}(\boldsymbol{\alpha})) + \min_{\mathbf{p}} C\sum_{i=1}^{N_h}(L_i(p_i) + \alpha_i p_i)$$
$$= -\max_{\mathbf{w}}(\mathbf{w}^T\mathbf{z}(\boldsymbol{\alpha}) - \Omega(\mathbf{w})) - \max_{\mathbf{p}} C\sum_{i=1}^{N_h}(-\alpha_i p_i - L_i(p_i))$$
$$= -\Omega^*(\mathbf{z}(\boldsymbol{\alpha})) - C\sum_{i=1}^{N_h} L_i^*(-\alpha_i). \quad (35)$$

Thus the dual problem is:

$$\max_{\boldsymbol{\alpha}\geq\mathbf{0}} -\Omega^*(\mathbf{z}(\boldsymbol{\alpha})) - C\sum_{i=1}^{N_h} L_i^*(-\alpha_i). \quad (36)$$

A.4 Computation of $\nabla^*\Omega(\mathbf{z}(\boldsymbol{\alpha}))$

In order to solve $\mathbf{w}$, we need to compute $\mathbf{w} = \nabla^*\Omega(\mathbf{z})$ given $\mathbf{z}$ and $\mathbf{w}_{h-1}$. Based on the conjugate dual property, we have the following problem:

$$\mathbf{w} = \arg\max_{\mathbf{w}} \mathbf{w}^T\mathbf{z} - \Omega(\mathbf{w})$$
$$= \arg\max_{\mathbf{w}} \mathbf{w}^T\mathbf{z} - \frac{\sigma}{2}\|\mathbf{w} - \mathbf{w}_{h-1}\|^2 - \frac{1}{2}(\sum_{k=1}^{K}\|\mathbf{w}_k - \mathbf{w}_k^{h-1}\|^2)$$
$$= \arg\max_{\mathbf{w}} -\frac{\sigma}{2}\|\mathbf{w} - \mathbf{w}_{h-1} - \frac{\mathbf{z}}{\sigma}\|^2 - \frac{1}{2}(\sum_{k=1}^{K}\|\mathbf{w}_k - \mathbf{w}_k^{h-1}\|^2)$$
$$= \arg\min_{\mathbf{w}} \frac{\sigma}{2}\|\mathbf{w} - \mathbf{w}_{h-1} - \frac{\mathbf{z}}{\sigma}\|^2 + \frac{1}{2}(\sum_{k=1}^{K}\|\mathbf{w}_k - \mathbf{w}_k^{h-1}\|^2) \quad (37)$$

Problem (37) is strictly convex problem, thus a unique minimizer exits, and can be computed in close-form. According to (Martins *et al.* 2011), we give the detailed solution as shown in Algorithm 3.

Algorithm 3 Computation of $\mathbf{w} = \nabla^*\Omega(\mathbf{z})$

Require: $\mathbf{z}, \mathbf{w}_{h-1}$, parameter $\frac{1}{\sigma}$.
Initialize $\boldsymbol{\omega} = \frac{z}{\boldsymbol{\sigma}}$.
Compute $\widehat{o}_k = \|\boldsymbol{\omega}_k\|$ where $\boldsymbol{\omega}_k$ is associated with $\mathbf{w}_k$ for $k = 1, \ldots, K$
Sort $\widehat{\mathbf{o}}$ to obtain $\bar{\mathbf{o}}$ such that $\bar{o}_1 \geq \cdots \geq \bar{o}_K$.
Find $\rho = \max\{k|\bar{o}_k - \frac{s}{1+ks}\sum_{i=1}^{k}\bar{o}_i > 0, k = 1, \ldots, K\}$.
Compute a threshold value $\zeta = \frac{s}{1+\rho s}\sum_{i=1}^{\rho}\bar{o}_i$.
Calculate $\mathbf{o}$, where $o_k = \begin{cases}\widehat{o}_k - \zeta, & \text{if } \widehat{o}_k > \zeta,\\ 0, & \text{Otherwise.}\end{cases}$
Calculate $\widehat{\boldsymbol{\omega}}_k = \begin{cases}\frac{o_k}{\|\boldsymbol{\omega}_k\|}\boldsymbol{\omega}_k, & \text{if } o_k > 0,\\ 0, & \text{Otherwise.}\end{cases}$
Let $\mathbf{w} = [\widehat{\boldsymbol{\omega}}_k]_{k=1}^{K}$ and return $\mathbf{w}$.

A.5 Online Update of Imbalance Measures

In this paper, we focus on three performance measures: F-measure, AUROC and AUPRC instead of mistake number or classification loss used in traditional methods. According to Algorithm 2, we need to maintain and update the performance measures M_{h+1}^j at each iteration h. However, if we directly compute M_{h+1}^j, it is computation expensive and requires to store all historical predictions and labels, which is really inefficient. Instead, we present how to update M_{h+1}^j by only using M_h and current $\mathbf{f_h}, \mathbf{y}_h$.

For F-measure, let $\bar{\mathbf{y}}_h = (\mathbf{y}_h + 1)/2$ and $\widehat{\mathbf{y}}_h = \text{sign}(\mathbf{f}_h > 0)$, $a_h = \sum_{\tau=1}^{h}\bar{\mathbf{y}}_\tau \cdot \widehat{\mathbf{y}}_\tau$, $c_h = \sum_{\tau=1}^{h}\sum\bar{\mathbf{y}}_\tau + \sum_{\tau=1}^{h}\sum\widehat{\mathbf{y}}_\tau$. We can calculate F-measure as: $F_{h+1} = \frac{2a_h}{c_h}$. In order to compute F-measure incrementally, we only need to update a_h and c_h as:

$$a_{h+1} = a_h + \bar{\mathbf{y}}_{h+1} \cdot \widehat{\mathbf{y}}_{h+1},$$
$$c_{h+1} = c_h + \sum\bar{\mathbf{y}}_{h+1} + \sum\hat{\mathbf{y}}_{h+1}.$$

AUROC and AUPRC are different from F-measure in that they need to compute the area value under various thresholds. We introduce two auxiliary hash table L_+^h and L_-^h of size m that partition $(0,1)$ into m uniform ranges. For $i \in \{1, \ldots, m\}$, $L_+^h[i]$ stores the number of positive examples before (including) h-th iteration with predictions f such that $\sigma(f) \in [(i-1)/m, i/m)$. Similarly, L_-^h stores negative examples with $\sigma(f) \in [(i-1)/m, i/m)$. σ is the sigmoid function that normalize f to $(0,1)$. Let N_h^+ and N_h^- denote the number of positive and negative examples respectively. We then compute the True Positive Rate (TPR) and False Positive Rate (FPR) as: $\text{TPR}(i) = \sum_{j=i+1}^{m} L_+^h[j]/N_h^+$ and $\text{FPR}(i) = \sum_{j=i+1}^{m} L_-^h[j]/N_h^-$. Thus,

$$\text{AUROC} = \frac{1}{2}\sum_{i=0}^{m-1}[\text{FPR}(i{+}1){-}\text{FPR}(i)][\text{TPR}(i){+}\text{TPR}(i{+}1)].$$

Similarly, Precision (P) and Recall (R) are computed as: $\text{P}(i) = \sum_{j=i+1}^{m} L_+^h[j]/\sum_{j=i+1}^{m}(L_h^+[j] + L_h^-[j])$ and $\text{R}(i) = \text{TPR}(i)$. Similarly,

$$\text{AUPRC} = \frac{1}{2}\sum_{i=0}^{m-1}[\text{R}(i) - \text{R}(i+1)][\text{P}(i) + \text{P}(i+1)].$$

In order to compute AUROC and AUPRC incrementally, we only need to maintain and update L_+^h and L_-^h.

A.6 Proof of Proposition 1

We mainly consider F1-score, whose computation is

$$F(h) = \frac{2(P_1 - \texttt{fn})}{2P_1 - \texttt{fn} + \texttt{fp}},$$

where h can be any hypothesis (classifiers, models, etc.), fn and fp denote the false negative probability and false positive probability, respectively.

Following (Parambath *et al.* 2014), we define the following notations for binary classification:

$$\boldsymbol{a}(\theta) = [1 - \frac{\theta}{2}, \frac{\theta}{2}] \in \mathbb{R}^2,$$
P_1 : the marginal probability of the positive instances,
$\boldsymbol{e} = [\texttt{fn}, \texttt{fp}] \in \mathbb{R}^2$: error profile,
$\boldsymbol{E}(h) = [\texttt{fn}, \texttt{fp}] \in \mathbb{R}^2$: error profile of h,
$$\boldsymbol{e}(\theta) \in \arg\min_{\boldsymbol{e}' \in \mathcal{E}}\langle \boldsymbol{a}(\theta), \boldsymbol{e}'\rangle.$$

Lemma 1. (Proposition 4 in (Parambath *et al.* 2014)) Let $F^* = \max_{\boldsymbol{e}' \in \mathcal{E}(\mathcal{H})} F(\boldsymbol{e}')$. We have:

$$\boldsymbol{e} \in \arg\min_{\boldsymbol{e}' \in \mathcal{E}(\mathcal{H})}\langle \boldsymbol{a}(F^*), \boldsymbol{e}'\rangle \Leftrightarrow F(\boldsymbol{e}) = F^*.$$

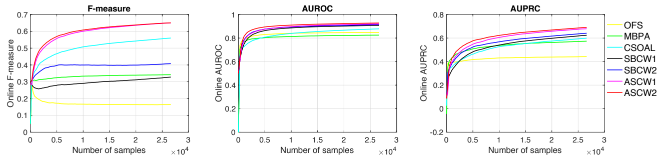
Lemma 2. (In the proof of Proposition 6 in (Parambath *et al.* 2014)) $F(\boldsymbol{e}) = t \Leftrightarrow \langle a(t), \boldsymbol{e}\rangle = \min_{\boldsymbol{e}' \in \mathcal{E}(\mathcal{H})}\langle \boldsymbol{a}(t), \boldsymbol{e}'\rangle = \frac{2P_1(t-1)}{2}$.

Here we re-present Proposition 1 as follows:

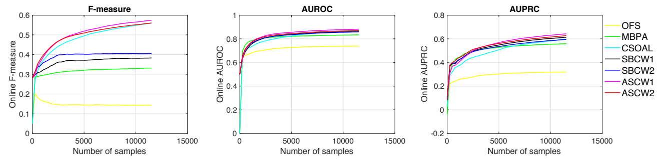
Proposition 1. Given the evenly distributed values $\theta_1, ..., \theta_K$ and the cost vector $\boldsymbol{a}(\theta) = [1 - \frac{\theta}{2}, \frac{\theta}{2}]$, let $\Delta = \frac{\theta_j - \theta_{j+1}}{2} = \frac{1}{2K}$. Denote $F^* = \max_{\boldsymbol{e}} F(\boldsymbol{e})$ the maximum F-measure and $F(\boldsymbol{\mu})$ a function of $\boldsymbol{\mu}$ that computes the F-measure achieved by $\boldsymbol{\mu}$. Assume that $\{\boldsymbol{\mu}_h^1, ..., \boldsymbol{\mu}_h^K\}$ minimizes the cost-sensitive loss to a certain degree and $\boldsymbol{E}(\boldsymbol{\mu}) = [\texttt{fn}, \texttt{fp}]$, i.e., false negative probability and false positive probability. Then the F-measure achieved by Algorithm 2 has the following lower bound as long as h increases:

$$\max_{j=1,\ldots,K} F(\boldsymbol{\mu}_h^j) \geq F^* - \Delta - \frac{\epsilon_0}{P_1},$$

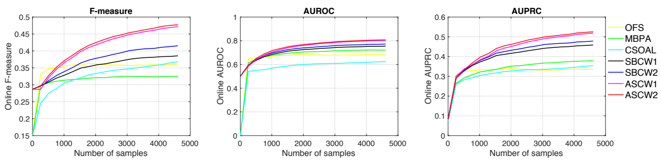
where $k = \arg\max_{j=1,\ldots,K} F(\boldsymbol{\mu}_h^j)$ and $\langle \boldsymbol{a}(\theta_j), \boldsymbol{E}(\boldsymbol{\mu}_h^k)\rangle \leq \min_{\boldsymbol{\mu}}\langle \boldsymbol{a}(\theta_j), \boldsymbol{E}(\boldsymbol{\mu})\rangle + \epsilon_0$.



(a) real-sim with ratio 1:5

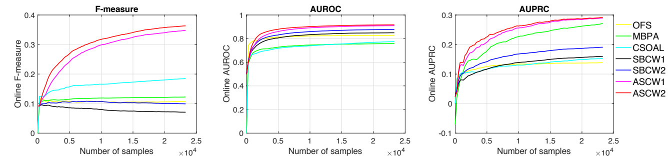


(b) rcv1 with ratio 1:5

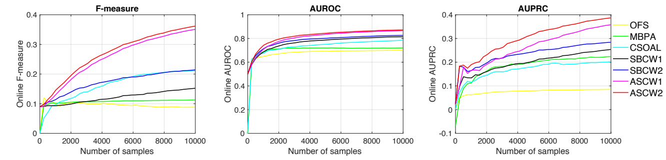


(c) news20 with ratio 1:5

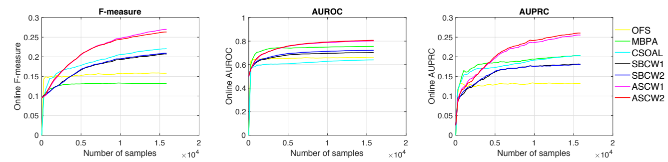
Figure S1: Online performance for ratio 1:5



(a) real-sim with ratio 1:20



(b) rcv1 with ratio 1:20



(c) news20 with ratio 1:19

Figure S2: Online performance for ration 1:20 (1:19)

Proof.

$$
\begin{aligned}
& F^* - F(\boldsymbol{\mu}_h^j) \\
=& F^* - \frac{2(P_1 - \boldsymbol{E}(\boldsymbol{\mu}_h^j))}{2P_1 - \boldsymbol{E}_1(\boldsymbol{\mu}_h^j) + \boldsymbol{E}_2(\boldsymbol{\mu}_h^j)} \\
=& \frac{2P_1(F^* - 1) + \langle \boldsymbol{a}(\theta^*), \boldsymbol{E}(\boldsymbol{\mu}_h^j) \rangle}{2P_1 - \boldsymbol{E}_1(\boldsymbol{\mu}_h^j) + \boldsymbol{E}_2(\boldsymbol{\mu}_h^j)} \\
=& \frac{\langle \boldsymbol{a}(\theta^*) - \boldsymbol{a}(\theta_j), \boldsymbol{E}(\boldsymbol{\mu}_h^j) \rangle + \langle \boldsymbol{a}(\theta_j), \boldsymbol{E}(\boldsymbol{\mu}_h^j) + 2P_1(\theta^* - 1) \rangle}{2P_1 - \boldsymbol{E}_1(\boldsymbol{\mu}_h^j) + \boldsymbol{E}_2(\boldsymbol{\mu}_h^j)} \\
=& \frac{(\theta^* - \theta_j) + \langle \boldsymbol{a}(\theta_j), \boldsymbol{E}(\boldsymbol{\mu}_h^j) \rangle + 2P_1(\theta^* - 1)}{2P_1 - \boldsymbol{E}_1(\boldsymbol{\mu}_h^j) + \boldsymbol{E}_2(\boldsymbol{\mu}_h^j)} \\
\leq& \frac{(\theta^* - \theta_j) + \langle \boldsymbol{a}(\theta_j), e(\theta_j) \rangle + \epsilon_0 + 2P_1(\theta^* - 1)}{2P_1 - \boldsymbol{E}_1(\boldsymbol{\mu}_h^j) + \boldsymbol{E}_2(\boldsymbol{\mu}_h^j)} \\
=& \theta^* - \theta_j + \frac{\epsilon_0}{2P_1 - \boldsymbol{E}_1(\boldsymbol{\mu}_h^j) + \boldsymbol{E}_2(\boldsymbol{\mu}_h^j)} \\
\leq& \Delta + \frac{\epsilon_0}{P_1}. \qquad (1)
\end{aligned}
$$

□

A.7 Additional Results

In Figure 1, we only show the online performance with respect to the ratio 1:10. Here we show other imbalance ratios in Figure S1 and Figure S2.

References

Andre Filipe Torres Martins, Noah Smith, Eric Xing, Pedro Aguiar, and Mario Figueiredo. Online learning of structured predictors with multiple kernels. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 507–515, 2011.

Shameem Puthiya Parambath, Nicolas Usunier, and Yves Grandvalet. Optimizing f-measures by cost-sensitive classification. In *Advances in Neural Information Processing Systems*, pages 2123–2131, 2014.

Maurice Sion. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

Mingkui Tan, Li Wang, and Ivor W Tsang. Learning sparse svm for feature selection on very high dimensional datasets. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 1047–1054, 2010.