CONE: A Cost-Sensitive Classification Wrapper for Iterative F-measure Optimization: Supplementary Material

Kevin Bascol
Laboratoire Hubert Curien UMR 5516,
Univ Lyon, UJM-Saint-Etienne
F-42023, Saint-Etienne, France
Bluecime inc., France

Guillaume Metzler
Laboratoire Hubert Curien UMR 5516,
Univ Lyon, UJM-Saint-Etienne
F-42023, Saint-Etienne, France
Blitz inc., France

Rémi Emonet, Amaury Habrard, Marc Sebban
Laboratoire Hubert Curien UMR 5516,
Univ Lyon, UJM-Saint-Etienne
F-42023, Saint-Etienne, France

Elisa Fromont
IRISA/Inria,
Univ. Rennes 1,
35042 Rennes cedex, France

The goal of this document is to:

- detail the proof of the results provided in the main article,
- develop a multi-class extension,
- provide exhaustive numerical values used to plot the curves (for easier reproducibility).

For the sake of clarity we will remind each statement before giving its proof. We also recall the notations and the definitions that are used for our purposes.

In the body of the paper $\mathbf{E}(h) = (e_1(h), e_2(h), e_3(h), e_4(h))$ has been simplified as $(e_1(h), e_2(h))$ as we have a redundancy with $e_3 = e_2$ (i.e. a FN of class 2 corresponds to a FP of class 1) and $e_4 = e_1$.

In the binary setting and using the previous notations, the F-Measure is defined by:

$$F(e) = \frac{(1+\beta^2)(P-e_1)}{(1+\beta^2)P - e_1 + e_2}. \tag{1}$$

1 Main results of the article

In this section, we provide all the proofs of the main article but only in the binary setting.

1.1 On Pseudo-linearity of F-Measure

We aim to prove the following proposition:

Proposition 1. *The F-measure, F, is a pseudo-linear function.*

Proof. We need to show that both F and $-F$ are pseudo-convex, i.e. we have:

$$\nabla F(\boldsymbol{e}), (\boldsymbol{e}' - \boldsymbol{e}) \geq 0 \implies F(\boldsymbol{e}') \geq F(\boldsymbol{e}). \tag{2}$$

The gradient of the F-measure is defined by:

$$\nabla F(\boldsymbol{e}) = -\frac{1+\beta^2}{((1+\beta^2)P - e_1 + e_2)^2} \begin{pmatrix} \beta^2 P + e_2 \\ P - e_1 \end{pmatrix}.$$

We now write the left hand side of the implication (2):

$$\begin{aligned}
\langle \nabla F(\boldsymbol{e}), (\boldsymbol{e}' - \boldsymbol{e}) \rangle &\geq \\
-\frac{1+\beta^2}{((1+\beta^2)P - e_1 + e_2)^2} \left[\beta^2 P(e_1' - e_1) + (P - e_1)(e_2' - e_2)\right] &\geq \\
-\beta^2 P(e_1' - e_1) - (P - e_1)(e_2' - e_2) &\geq \\
\beta^2 P(e_1 - e_1') + P(e_2 - e_2') + e_1 e_2' - e_1' e_2 &\geq \\
-\beta^2 P e_1' + \beta^2 P e_1 + P e_2 - P e_2' + e_1 e_2' - e_1' e_2 &\geq
\end{aligned}$$

We now add $-P(e_1 + e_2)$ on both side of the inequality, so we have:

$$-(1+\beta^2)Pe_1' - Pe_1 + Pe_2 - e_1'e_2 \quad \geq \quad -(1+\beta^2)Pe_1 - Pe_1' + P$$

Then by adding $(1+\beta^2)P^2$ on both sides of the inequality we get:

$$\begin{aligned}
(1+\beta^2)P(P - e_1') - (P - e_1')e_1 + (P - e_1')e_2 &\geq (1+\beta^2)P(P - \\
(P - e_1')((1+\beta^2)P - e_1 + e_2) &\geq (P - e_1)((1 + \\
(1+\beta^2)(P - e_1')((1+\beta^2)P - e_1 + e_2) &\geq (1+\beta^2)(P - \\
F(\boldsymbol{e}') &\geq F(\boldsymbol{e}).
\end{aligned}$$

The proof is similar for $-F$. We have shown that both F and $-F$ are pseudo-convex so F is pseudo-linear.

□

1.2 An optimal value of M'

Now, we would like to give an explicit value for M' that can be obtained by solving the following optimization problem:

$$\max_{\boldsymbol{e}' \in \mathcal{E}(\mathcal{H})} e'_2 - e'_1 \quad s.t. \quad F_\beta(\boldsymbol{e'}) > F_\beta(\boldsymbol{e}).$$

In the binary case, setting $\boldsymbol{e} = (e_1, e_2)$ and $\boldsymbol{e'} = (e'_1, e'_2)$. We can write $F_\beta(\boldsymbol{e'}) > F_\beta(\boldsymbol{e})$ as:

$$\frac{(1+\beta^2)(P-e'_1)}{(1+\beta^2)P - e'_1 + e'_2} > \frac{(1+\beta^2)(P-e_1)}{(1+\beta^2)P - e_1 + e_2},$$

Now we develop and reduce these expressions.

$$\begin{aligned}
(P-e_1)[(1+\beta^2)P - e'_1 + e'_2] &> (P-e'_1)[(1+\beta^2)P - e_1 + e_2],\\
-(1+\beta^2)Pe'_1 + (P-e'_1)(e_2-e_1) &> -(1+\beta^2)Pe_1 + (P-e_1)(e'_2-e'_1),\\
(1+\beta^2)P(e_1-e'_1) + P(e_2-e_1+e'_1-e'_2) &> e_2e'_1 - e_1e'_2.
\end{aligned}$$

Now, we set: $e'_1 = e_1 + \alpha_1$ and $e'_2 = e_2 + \alpha_2$. In other words, we study how much we have to change $\boldsymbol{e}'$ from $\boldsymbol{e}$ to solve our problem. We can then write:

$$\begin{aligned}
-(1+\beta^2)P\alpha_1 + P(\alpha_1 - \alpha 2) &> e_2(e_1+\alpha_1) - e_1(e_2+\alpha_2),\\
\alpha_1(-(1+\beta^2)P + P - e_2) + \alpha_2(-P+e_1) &> 0,\\
\alpha_1(\beta^2P + e_2) &< -\alpha_2(P-e_1).
\end{aligned}$$

Thus we have the following constraints on α_1, α_2 :

$$\begin{aligned}
\alpha_1 &\in [-e_1, P-e_1],\\
\alpha_2 &\in [-e_2, N-e_2],\\
\alpha_1 &< \frac{-\alpha_2(P-e_1)}{\beta^2P+e_2},
\end{aligned}$$

and the optimization problem can be rewritten as:

$$\begin{aligned}
\max_{\boldsymbol{\alpha}} \quad & \alpha_2 - \alpha_1,\\
s.t. \quad & \alpha_1 < \frac{-\alpha_2(P-e_1)}{\beta^2P+e_2},\\
& \alpha_1 \in [-e_1, P-e_1],\\
& \alpha_2 \in [-e_2, N-e_2].
\end{aligned}$$

introduire la ou les figure(s) pour expliquer les differents cas possibles, a faire avec Inkscape ou autre, expliquer les configuarations. dire que l'analyse de Mmin ou max est identique, donc on ne le fait que pour Mmax en entier, puis on synthétise pour Mmin

$$\alpha_1 = \frac{-\alpha_2(P-e_1)}{\beta^2P+e_2}. \tag{3}$$

We are looking for the set of values of α_2 such that α_1 belongs to $[-e_1, P-e_1]$ using Equation (3). We get:

$$\left[-(\beta^2P+e_2), e_1\frac{\beta^2P+e_2}{P-e_1}\right].$$

Thus, the set of admissible values for α_2 is

$$\left[-e_2, \min\left(e_1\frac{\beta^2P+e_2}{P-e_1}, N-e_2\right)\right].$$

The norm $\|e'\|_2^2$ reaches its maximum at the limit of the predefined set, i.e., when $\alpha_2 = -e_2$ or $\alpha_2 = \min\left(e_1\frac{\beta^2P+e_2}{P-e_1}, N-e_2\right)$. We have the corresponding values of α_1 using (3).

Our tighter slope can now be derived from the computation of M', i.e., the value of $e'_2 - e'_1 = e_2 + \alpha_2 - (e_1 + \alpha_1)$.

2 The multi-class setting

For a given hypothesis $h \in \mathcal{H}$ learned from $\mathbf{X}$, the errors that h makes can be summarized in an error profile defined as $\mathbf{E}(h) \in \mathbb{R}^{2L}$:

$$\mathbf{E}(h) = (\mathtt{FN}_1(h), \mathtt{FP}_1(h), ..., \mathtt{FN}_L(h), \mathtt{FP}_L(h)),$$

where $\mathtt{FN}_i(h)$ (resp. $\mathtt{FP}_i(h)$) is the proportion of False Negative (resp. False Positive) that h yields for class i.

In a multiclass setting with L classes P_k, $k = 1, ..., L$ denotes the proportion of examples in class k and $\boldsymbol{e} = (e_1, e_2, ..., e_{2L-1}, e_{2L})$ denotes the proportions of misclassified examples composing the error profile.

The multiclass-micro F-Measure, $mcF_\beta(\boldsymbol{e})$ with L classes is defined by:

$$mcF_\beta(\boldsymbol{e}) = \frac{(1+\beta^2)(1 - P_1 - \sum_{k=2}^{L} e_{2k-1})}{(1+\beta^2)(1-P_1) - \sum_{k=2}^{L} e_{2k-1} + e_1}. \tag{4}$$

Moreover, the corresponding function a that assigns the misclassification costs is shown by (**?**) to be defined as, for all $t \in [0,1]$:

$$a(t) = \begin{cases} 1 + \beta^2 - t & for\ e_{2k-1},\ k = 2, ..., L\\ t & for\ e_1.\\ 0 & otherwise. \end{cases}$$

The corresponding value of Φ is $\frac{1}{\beta^2 \sum_{k=1}^{L} P_k}$.

3 Extended Experiments