CAp 2018:
CONE: A Cost-Sensitive Classification Wrapper for Iterative F-Measure Optimization

Kevin Bascol, Rémi Emonet, Elisa Fromont, Amaury Habrard,
Guillaume Metzler and Marc Sebban

22 juin 2018

Introduction

$$\textbf{Accuracy } A = \frac{TP + TN}{P + N}, \qquad \textbf{F-Measure } F = \frac{2(P - FN)}{2P - FN + FP}.$$

Let us now working in an imbalanced setting ($P << N$):

Classical classifiers, based on the minimization of the error rate, tend to predict the majority class $\Rightarrow A \simeq 1$.

However $\Rightarrow FN \simeq P \Rightarrow F = 0$. Accuracy is not suitable measure in an imbalanced setting compare to the F-Measure.

Problem F-Measure is non-convex $\rightarrow$ hard to optimize

$\Rightarrow$ Approximate F-Measure optimization with weighted accuracy.

The Idea from Parambath et al. (NIPS 2014)

- A weighting function $a(t) = \big(1 + \beta^2 - t, t\big)$, weights on FP and FN.
- A classifier $h \in \mathcal{H}$ and its error profile $\boldsymbol{e}(t) = (e_1(t), e_2(t))$

$\rightarrow$ Idea: find a link between the weighted error and the F-Measure, using the notion of *Pseudo-linearity*.

Base Result (from Parambath et al .2014)

$\rightarrow$ Propose an upper bound on the optimal F-Measure

Let $\varepsilon_0 \geq 0$ and $\varepsilon_1 \geq 0$, and assume that there exists $\Phi > 0$ such that for all $\boldsymbol{e}, \boldsymbol{e'}$ satisfying $F(\boldsymbol{e'}) > F(\boldsymbol{e})$, we have:

$$F(\boldsymbol{e'}) - F(\boldsymbol{e}) \leq \Phi \langle a(F(\boldsymbol{e'})), \boldsymbol{e} - \boldsymbol{e'} \rangle.$$

Then , let us take $\boldsymbol{e}^\star \in argmax\ F(\boldsymbol{e'})$ and denote $\boldsymbol{a}^\star = a(F(\boldsymbol{e}^\star))$. Let furthermore $\boldsymbol{g} \in \mathbb{R}^d_+$ and $h \in \mathcal{H}$ satisfying the following two conditions:

$$(i)\ \|\boldsymbol{g} - \boldsymbol{a}^\star\|_2 \leq \varepsilon_0, \quad (ii)\ \langle \boldsymbol{g}, \mathbf{E}(h) \rangle \leq \min_{\boldsymbol{e'} \in \mathcal{E}(\mathcal{H})} \langle \boldsymbol{g}, \boldsymbol{e'} \rangle + \varepsilon_1.$$

We have:

$$F(\mathbf{E}(h)) \geq F(\boldsymbol{e}^\star) - \Phi(2\varepsilon_0 M + \varepsilon_1), \quad M = \max_{\boldsymbol{e'} \in \mathcal{E}(\mathcal{H})} \|\boldsymbol{e'}\|_2,$$

where $F(\boldsymbol{e}^\star)$ is the optimal value of the F-Measure.

Geometric Interpretation

A weighting function:

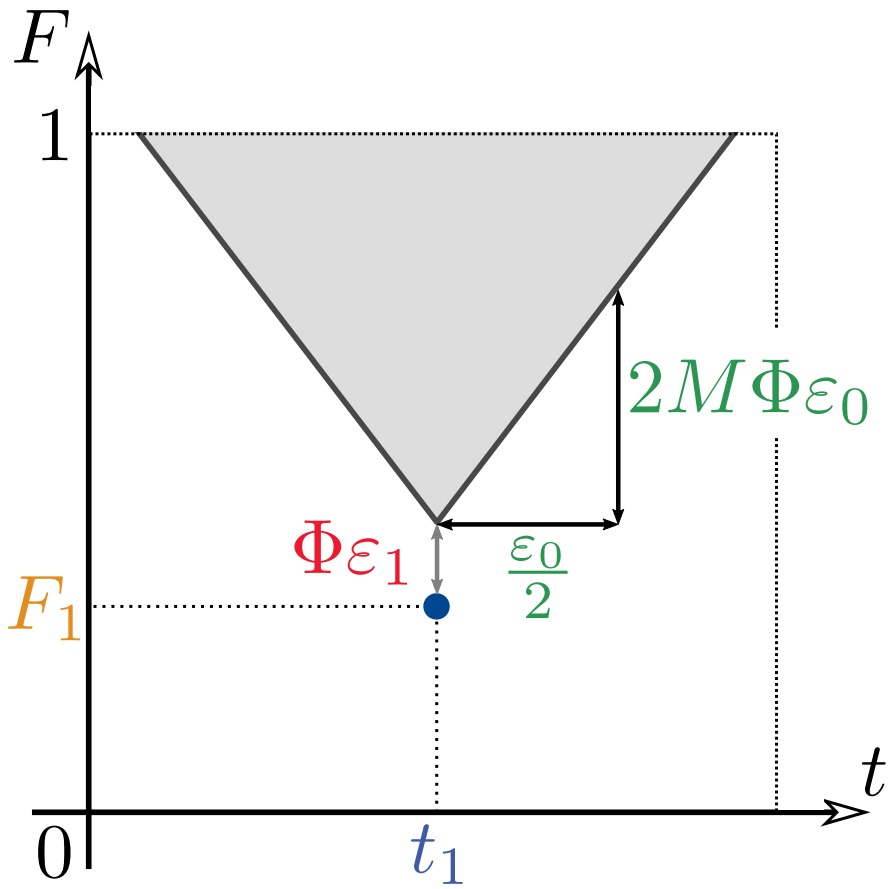
$$a(t) = (1 + \beta^2 - t, t).$$

We can rewrite the point

$\|\boldsymbol{g} - \boldsymbol{a}^\star\|_2 \leq \varepsilon_0$, as
$\|a(t_1) - a(t)\|_2 \leq 2\|t_1 - t\|_2 = \varepsilon_0$,

and the bound in function of t:

$$\begin{aligned} F(\boldsymbol{e}(t)) \quad \leq \quad & F(\boldsymbol{e}(t_1)) \\ & + 4\Phi M \|t_1 - t\|_2 \\ & + \Phi\varepsilon_1. \end{aligned}$$



A Tighter Slope

→ Use $\sqrt{2}$ as a Lipschitz constant of a, find a value of M.

Considering the assumptions of the base result, for all $\boldsymbol{e} \in \mathcal{E}(\mathcal{H})$ and all t we have:

$$F(\boldsymbol{e}(t)) \leq F(\boldsymbol{e}(t_1)) + \Phi\sqrt{2}(\|\boldsymbol{e}\|_2 + M')\|t_1 - t\|_2 + \Phi\varepsilon_1.$$

In other words, we refined the slope of the cones to $\sqrt{2}\Phi\,(\|\boldsymbol{e}\|_2 + M')$.

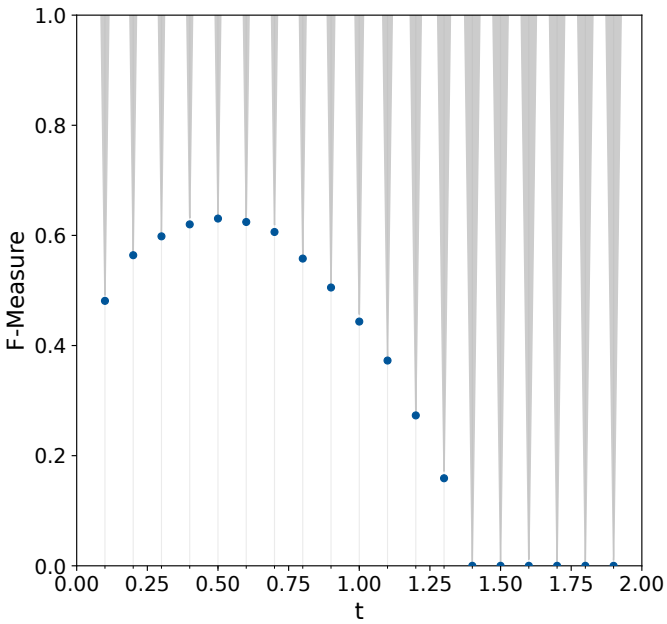
$$M' = \max_{\boldsymbol{e'} \in \mathcal{E}(\mathcal{H})} \|\boldsymbol{e'}\|_2, \quad s.t.\ F_\beta(\boldsymbol{e'}) > F_\beta(\boldsymbol{e}).$$

Writing $\boldsymbol{e'} = \boldsymbol{e} + \boldsymbol{\alpha}$, finding M' consist of maximizing a convex function on a square with a linear constraint.

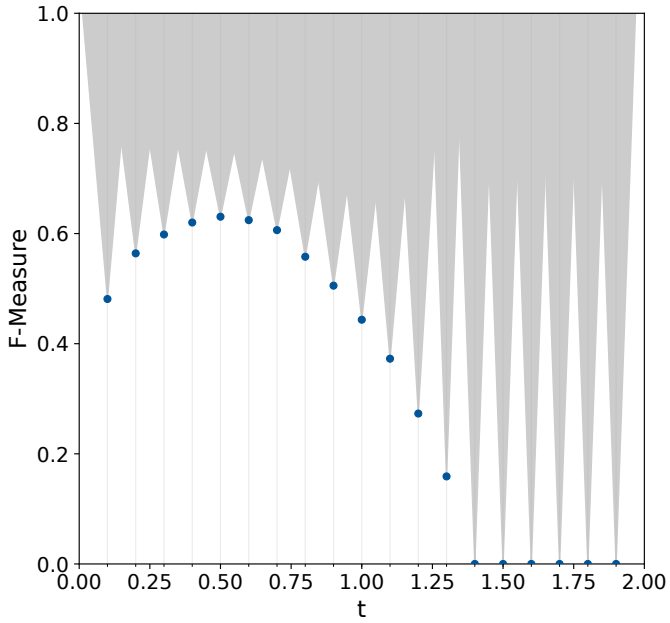
→ Need to look what happen on the border of the square.

A Tighter Slope ?

Unreachable region obtained with the bounds on points from a grid



Parambath et al.



With a tighter slope

We observe that $F(\boldsymbol{e}(t)) \simeq 0$ when t is large.

$\rightarrow$ Can we reduce the space of research ?

Search Space Pruning

$\rightarrow$ Assumption : the learned classifiers are *optimal* $\Rightarrow$ $\varepsilon_1 = 0$.

$\rightarrow$ We can show that $(e_1 - e_2)(t) = (FN - FP)(t)$ is increasing.

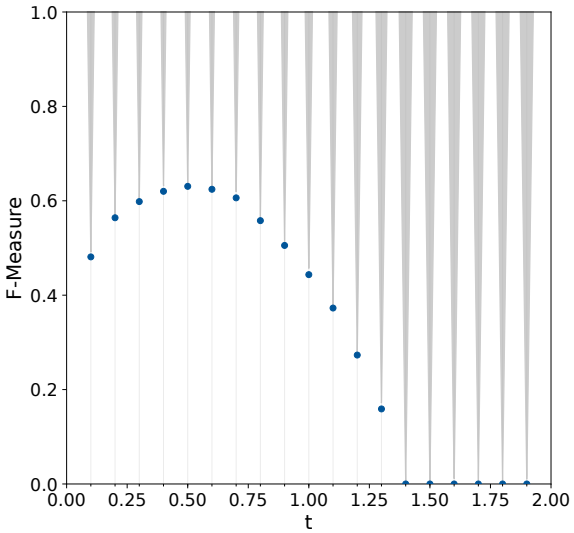
A bound on F-Measure

Let $t > t_1$, $\boldsymbol{e}(t)$ and $\boldsymbol{e}(t_1)$ the vector of misclassified examples obtained with an optimal classifier trained with costs $a(t)$ and $a(t_1)$ respectively. We have:

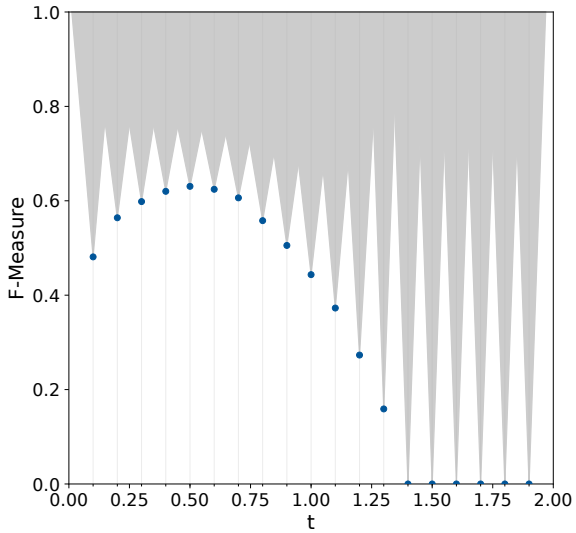
$$F_\beta(\boldsymbol{e}(t)) \leq (1+\beta^2)\frac{\frac{1+\beta^2}{t_1}TP(t_1)}{\beta^2\frac{1+\beta^2}{t_1}TP(t_1) + P}.$$

Illustration of the Pruning Effect

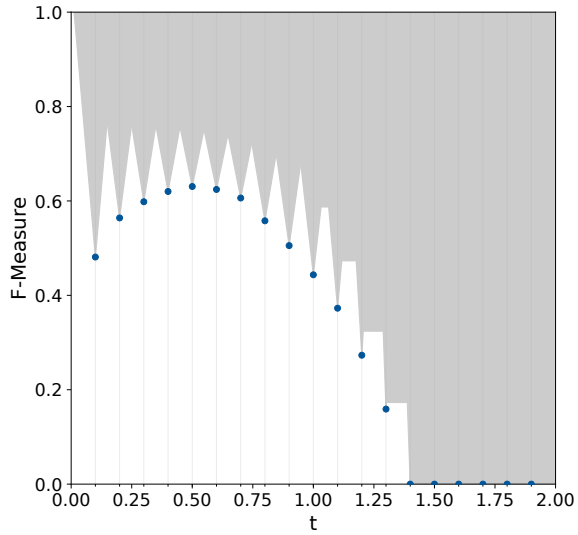
Unreachable region obtained with the bounds on points from a grid



Parambath et al.



CONE



CONE+ pruning

Presentation of **CONE**

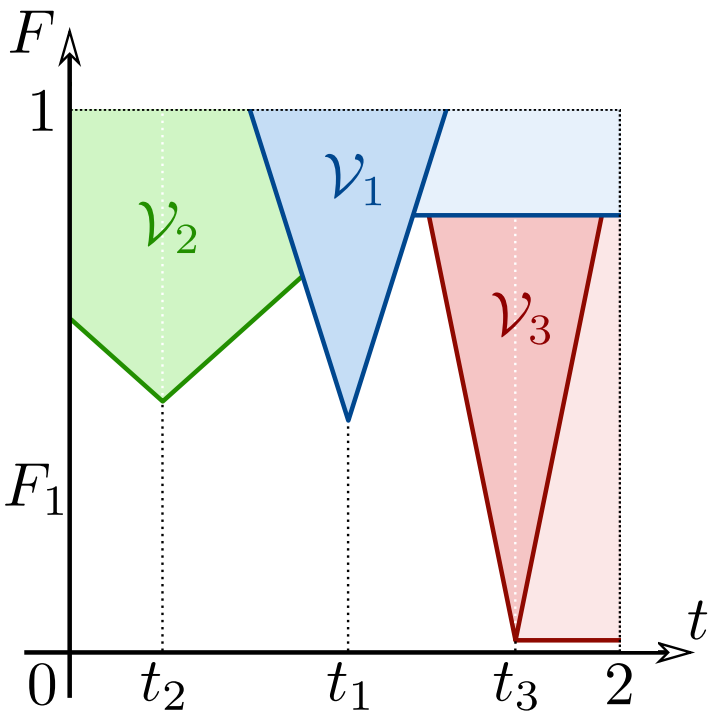


Illustration of **CONE** on the three first iterations.

$\boldsymbol{\nu}_1$: First cone with t in the middle of the search space: $t_1 = 1$

$\rightarrow$ Highest remaining $F = 1$ for $t \in [0,\ 0.6]$

$\boldsymbol{\nu}_2$: Next cone with t in the middle of this interval: $t_2 = 0.3$

$\rightarrow$ Highest remaining $F = 0.7$ for $t \in [1.3,\ 2]$

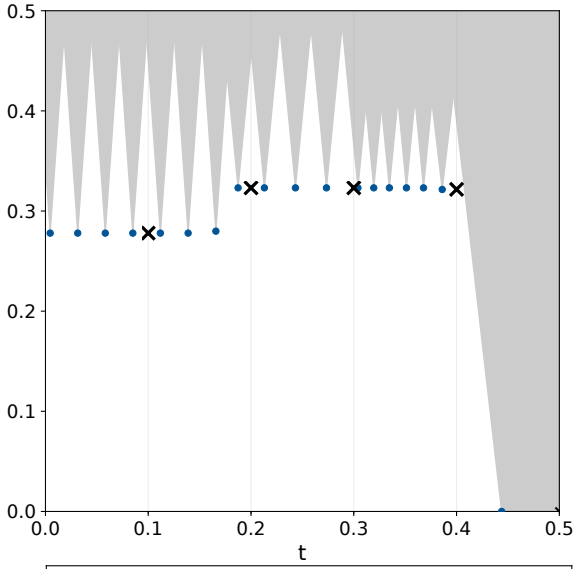
$\boldsymbol{\nu}_3$: Next cone with t in the middle of this interval: $t_3 = 1.65$

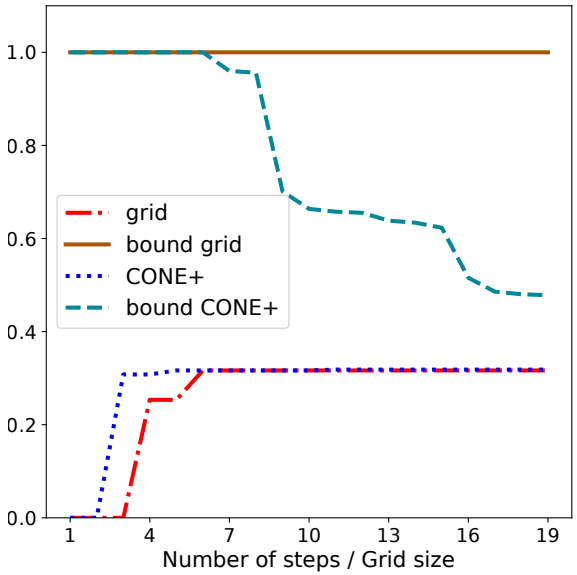
$\rightarrow$ Highest remaining $F = 0.7$ for $t \in [1.3,\ 1.35]$

$\boldsymbol{\nu}_\infty$: Until we reach the best F possible

Experimental Results

Abalone 10

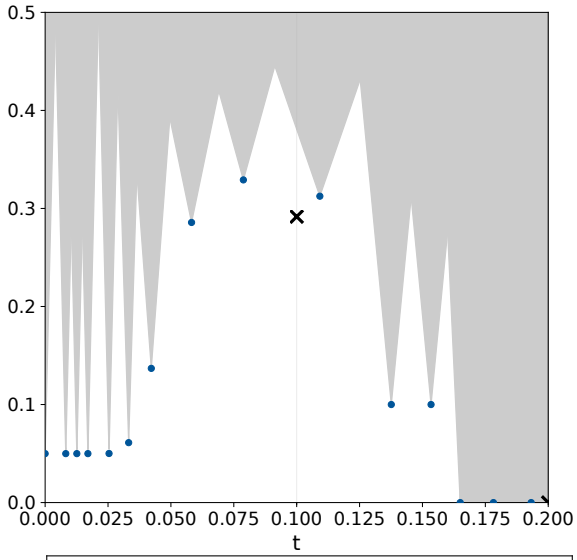


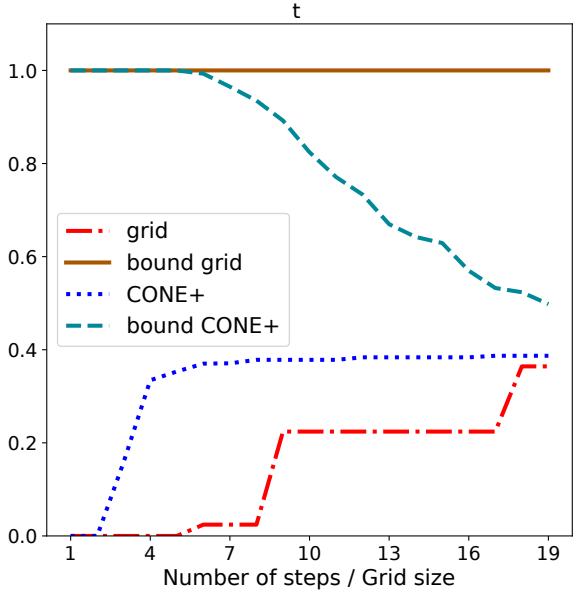


Examples of run of **CONE** (blue points and gray area) compared to Grid (black crosses)

Corresponding convergence of best F-measure and its bound in function of the number of classifier used

Yeast





Results Table

When we limit the grid size/number of cone to...

- ... 9 SVMs:

Dataset	P (%)	SVM*	SVM$_C$	SVM$_C^+$
Adult	23.9	66.1 (0.1)	66.5 (0.1)	66.5 (0.1)
Abalone10	15.2	30.2 (2.5)	31.0 (1.1)	32.3 (1.2)
IJCNN'01	9.6	61.6 (0.4)	61.0 (0.6)	61.6 (0.6)
Abalone12	6.4	16.1 (3.5)	12.2 (7.0)	17.0 (3.5)
Yeast	3.4	24.5 (16.3)	34.8 (8.3)	32.3 (12.2)
Wine	3.3	11.7 (11.3)	11.3 (10.8)	19.4 (6.6)

- ... 4 SVMs:

Dataset	P (%)	SVM*	SVM$_C$	SVM$_C^+$
Adult	23.9	65.8 (0.3)	66.5 (0.03)	66.5 (0.04)
Abalone10	15.2	30.7 (2.8)	12.2 (14.5)	30.8 (1.1)
IJCNN'01	9.6	61.0 (0.5)	61.0 (0.6)	61.0 (0.6)
Abalone12	6.4	0.0 (0.0)	0.0 (0.0)	15.9 (3.7)
Yeast	3.4	33.0 (18.0)	14.7 (12.0)	35.0 (8.4)
Wine	3.3	0.0 (0.0)	0.0 (0.0)	17.7 (4.4)

I.R.: Imbalance Ratio

SVM*: Reproduction of Parambath et al. algorithm

SVM$_C$: **CONE**

SVM$_C^+$: **CONE** with pruning

Conclusion & Future work

In this work,

- we derive a tighter bound,
- we propose an algorithm which optimize the search space,
- we show quick convergence results

In a possible future,
From a theoretical point of view:

- Search Space Pruning for small values of t
- Derive Generalization Guarantee over the F-Measure

Further experiments:

- Apply with other algorithms such as Neural Networks
- Dealing with the notion of sub-optimality in non-convex cases.

Thank you for your attention.