From Cost-Sensitive Classification to Tight F-measure Bounds

K. Bascol[1,3], R.Emonet[1], E. Fromont[2], A.Habrard [1], G. Metzler[1,4], and M. Sebban[1]

1. Univ. Lyon, UJM-Saint-Etienne, Laboratoire Hubert Curien UMR CNRS 5516, F-42023, Saint-Etienne, France
2. Univ. Rennes 1, IRISA/Inria, 35042 Rennes cedex, France
3. BLUECIME inc., France and 4. BLITZ inc., France

Abstract

In an **imbalanced** setting:

$\longrightarrow$ optimizing the classical accuracy tends to predict only the majority class;

$\longrightarrow$ optimizing imbalance-proof measures (as the F-Measure) is a tough task due to its non-convexity;

$\Longrightarrow$ approximate F-measure optimization by cost-sensitive approach.

We propose to:

- Write the difference of F-measures between the errors made by two hypotheses.
- Give an upper bound on the optimal reachable F-measure given the error made by the classifier and the used cost sensitive parameters.
- $\mathbf{CONE}$, an algorithm to iteratively optimize the F-measure.

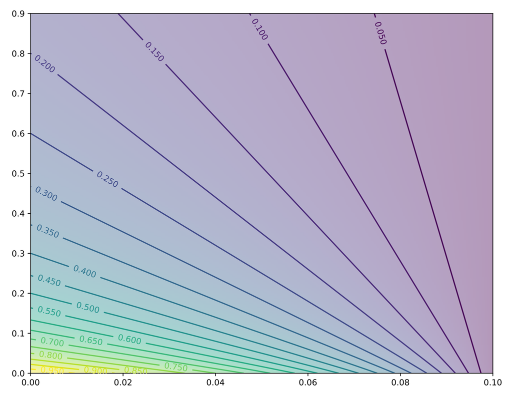
Notations and Base Result

Binary Classification

- $\boldsymbol{e} = (e_1, e_2) = (FN, FP)$ the error profile obtained from h.
- A function to assign costs on each class $\boldsymbol{a}(t) = (1+\beta^2 - t, t)$.
- $F(\boldsymbol{e}) = \dfrac{(1+\beta^2)(P-e_1)}{(1+\beta^2)P - e_1 + e_2}$ the associated F-measure.

Property F-measure

The level sets of the F-measure are hyperplanes: given $t \in [0,1]$, $F(\boldsymbol{e}) = t$ if and only if $\exists \boldsymbol{a}, b$, two functions such that $\langle \boldsymbol{a}(t), \boldsymbol{e}\rangle + b(t) = 0$.



A bound on the F-measure

Step 1: impact of a change in the error profile. Given two error profiles $\boldsymbol{e}$ and $\boldsymbol{e}'$ and the previous property of the F-measure:

$$\begin{aligned}\langle \boldsymbol{a}(F(\boldsymbol{e}')), \boldsymbol{e}-\boldsymbol{e}'\rangle &= \langle \boldsymbol{a}(F(\boldsymbol{e}')), \boldsymbol{e}\rangle + b(F(\boldsymbol{e}')),\\ &= (F(\boldsymbol{e}') - F(\boldsymbol{e}))\cdot\left((1+\beta^2)P_1 - e_1 + e_2\right),\end{aligned}$$

which leads to:

$$F(\boldsymbol{e}') - F(\boldsymbol{e}) = \Phi_{\boldsymbol{e}} \cdot \langle \boldsymbol{a}(F(\boldsymbol{e}')), \boldsymbol{e}-\boldsymbol{e}'\rangle, \tag{1}$$

Step 2: bounding the difference of F-measures. Suppose that a classifier trained with $\boldsymbol{a}(t)$ leads to $\boldsymbol{e}$ and $F(\boldsymbol{e})$ and consider $\boldsymbol{e}'$ obtained from an hypothetical classifier learned with $\boldsymbol{a}(t')$. Then, from Eq;(1), we have:

$$\begin{aligned}F(\boldsymbol{e}') - F(\boldsymbol{e}) &= \Phi_{\boldsymbol{e}}\left(\langle \boldsymbol{a}(t'), \boldsymbol{e}\rangle - \langle \boldsymbol{a}(t'), \boldsymbol{e}'\rangle\right),\\ &\le \Phi_{\boldsymbol{e}}\left(\langle \boldsymbol{a}(t), \boldsymbol{e}'\rangle + \varepsilon_1 - \langle \boldsymbol{a}(t'), \boldsymbol{e}'\rangle + (t'-t)(e_2-e_1)\right),\\ &\le \Phi_{\boldsymbol{e}}\varepsilon_1 + \Phi_{\boldsymbol{e}}\cdot(e_2 - e_1 - (e_2' - e_1'))(t'-t),\end{aligned}$$

where ε_1 : *sub-optimality* of the learned classifier w.r.t. the 0-1 loss

$$\langle \boldsymbol{a}(t), \boldsymbol{e}\rangle \le \varepsilon_1 + \min_{\boldsymbol{e}' \in \mathcal{E}(\mathcal{H})} \langle \boldsymbol{a}(t), \boldsymbol{e}'\rangle$$

$\longrightarrow \boldsymbol{e}' = (e_1', e_2')$ is unknown $\rightarrow$ bound it such that $F(\boldsymbol{e}') > F(\boldsymbol{e})$.

CONE: a Bound Driven Search Algorithm

Proposition. Let $\boldsymbol{e}$ be the error profile obtained with a classifier trained with the parameter t, $F(\boldsymbol{e})$ its associated F-measure value, $\Phi_{\boldsymbol{e}}$ as defined in Eq. (1), and $\varepsilon_1 > 0$ the sub-optimality of our linear classifier.

Then for all $t' < t$:

$$F(\boldsymbol{e}') \le F(\boldsymbol{e}) + \Phi_{\boldsymbol{e}}\varepsilon_1 + \Phi_{\boldsymbol{e}}\cdot(e_2 - e_1 - M_{max})(t'-t),$$

where $M_{max} = \underset{\substack{\boldsymbol{e}'' \in \mathcal{E}(\mathcal{H})\\ s.t.\ F(\boldsymbol{e}'') > F(\boldsymbol{e})}}{max} (e_2'' - e_1'')$

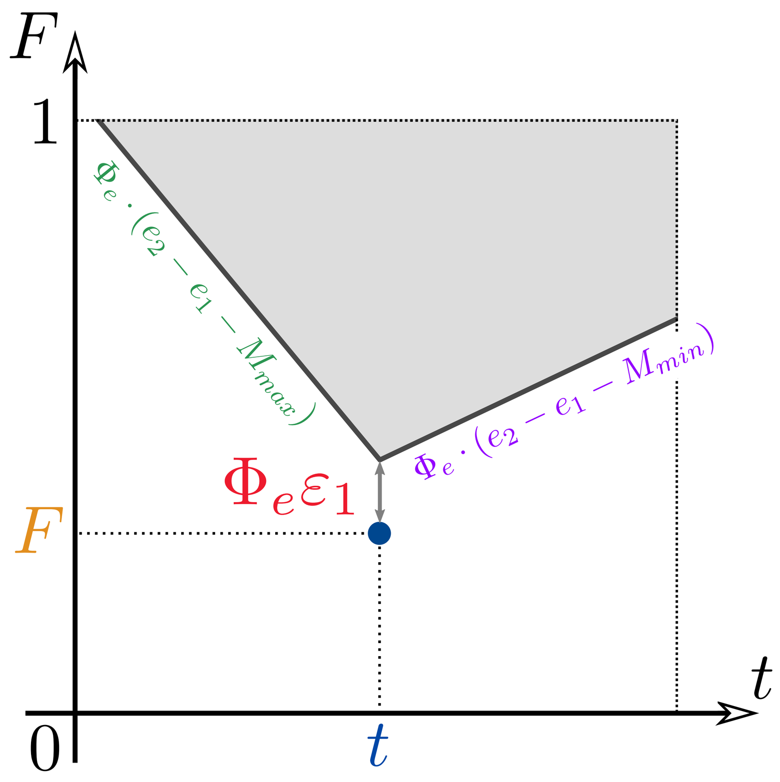
and, for all $t' > t$:

$$F(\boldsymbol{e}') \le F(\boldsymbol{e}) + \Phi_{\boldsymbol{e}}\varepsilon_1 + \Phi_{\boldsymbol{e}}\cdot(e_2 - e_1 - M_{min})(t'-t),$$

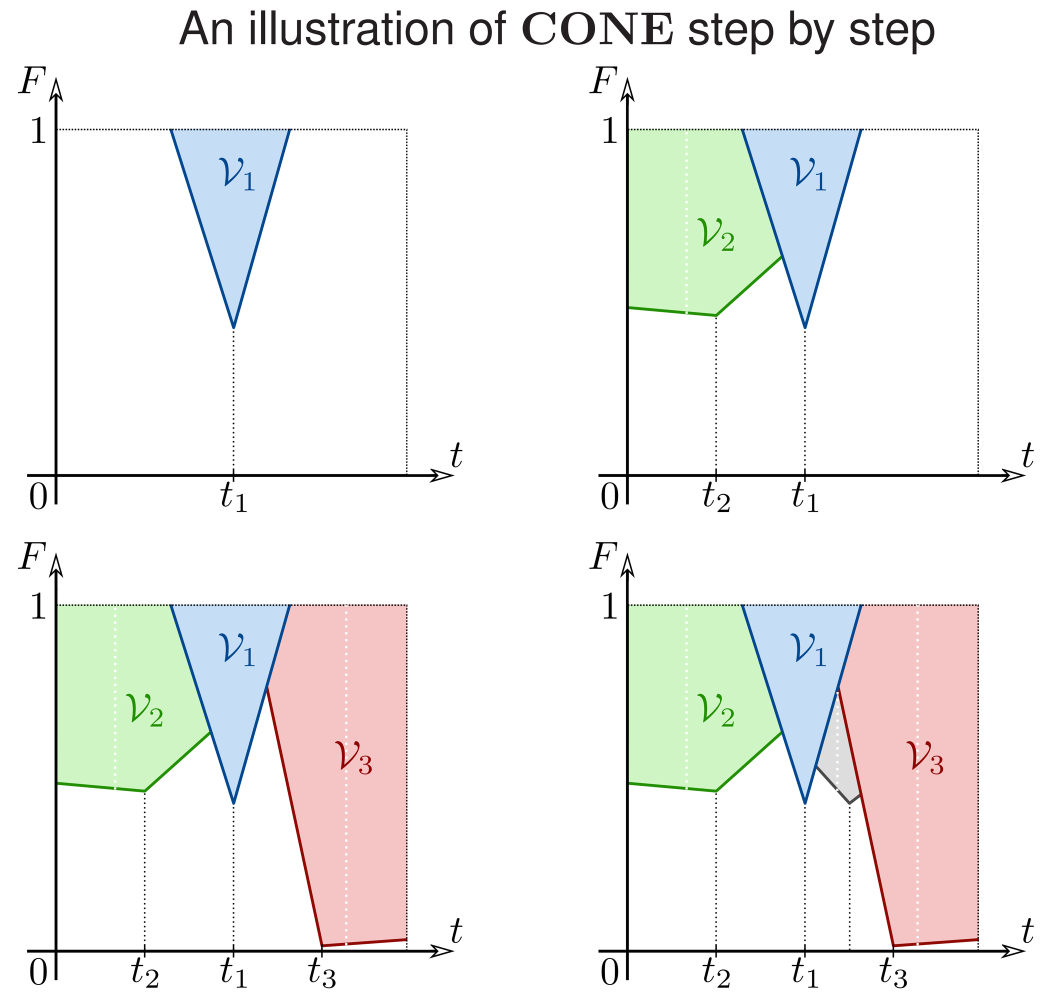
where $M_{min} = \underset{\substack{\boldsymbol{e}'' \in \mathcal{E}(\mathcal{H})\\ s.t.\ F(\boldsymbol{e}'') > F(\boldsymbol{e})}}{min} (e_2'' - e_1'')$.

A Geometric Interpretation

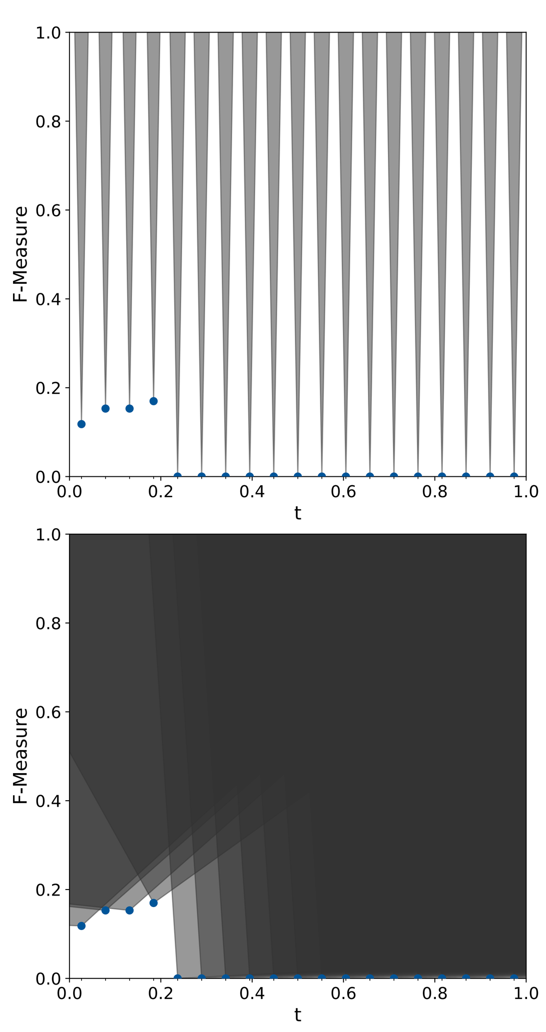
Our bound on $F(\boldsymbol{e}')$ can been seen the unreachable region of F-measure in the (t, F)−space.



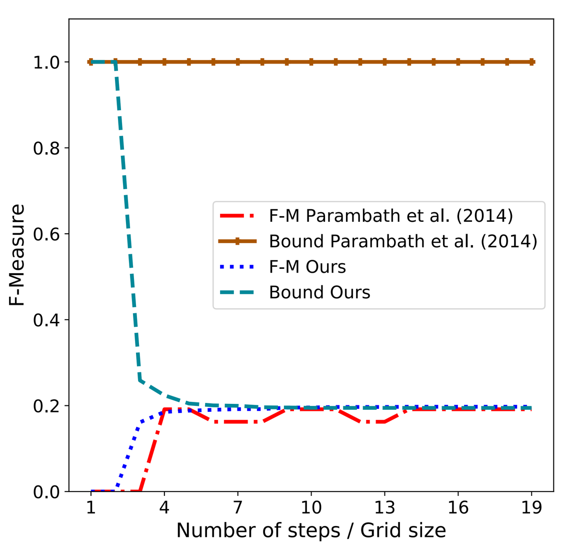
An illustration of $\mathbf{CONE}$ step by step

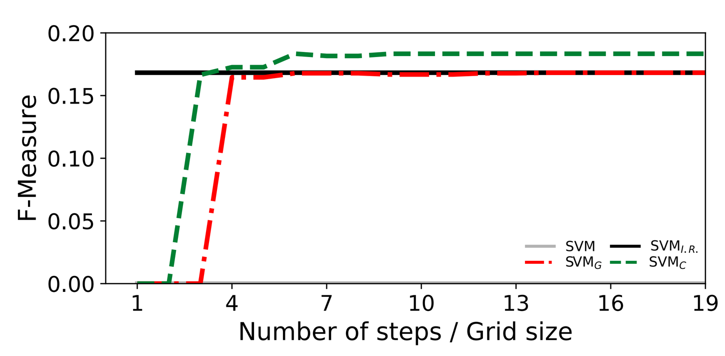


Practical Evaluation of Theoretical Guarantees



Examples of runs on Abalone12
top: wrapper from [1]; bottom: **CONE**;
both with SVM classifier ($C = 1$).





On Abalone12, in function of the number of **CONE** steps and [1] grid size:
left: on train set, evolution of the F-Measure and of the considered bound
right: on test set, evolution of the F-Measure

Results

- A more informative bound on the optimal F-measure than Parambath et al. [1].
- A faster convergence than Parambath et al. [1].
- Better results in average with a small number of iterations compared to [1,2,3].

Dataset	SVM	SVM$_{I.R.}$	SVM$_G$	SVM$_C$	SVM$_C^T$	LR^T	LR$^T_{I.R.}$	LR^T_G	LR$_B$
Adult	62.5 (0.2)	64.9 (0.3)	66.4 (0.1)	66.5 (0.1)	66.4 (0.1)	66.5 (0.1)	66.5 (0.1)	66.5 (0.1)	**66.6** (0.1)
Abalone10	0.0 (0.0)	30.9 (1.2)	**32.4** (1.3)	32.2 (0.8)	31.8 (1.9)	30.8 (2.2)	30.7 (1.9)	30.7 (1.9)	31.6 (0.6)
Satimage	0.0 (0.0)	23.4 (4.3)	20.4 (5.3)	20.6 (5.6)	**30.9** (2.0)	21.2 (11.1)	28.6 (1.9)	28.6 (1.9)	21.4 (4.6)
IJCNN	44.5 (0.4)	53.3 (0.4)	61.6 (0.6)	61.6 (0.6)	**62.6** (0.4)	59.4 (0.5)	56.5 (0.3)	56.5 (0.3)	59.2 (0.3)
Abalone12	0.0 (0.0)	16.8 (2.7)	16.8 (4.2)	**18.3** (3.3)	16.3 (3.0)	15.5 (3.1)	17.0 (3.3)	17.0 (3.3)	17.7 (3.7)
Pageblocks	48.1 (5.8)	39.6 (4.7)	66.4 (3.2)	62.8 (3.9)	**67.6** (4.0)	59.2 (8.1)	55.9 (6.4)	55.9 (6.4)	55.7 (5.7)
Yeast	0.0 (0.0)	29.4 (2.9)	38.6 (7.1)	39.0 (7.5)	35.4 (15.6)	37.4 (10.1)	**39.9** (6.5)	27.6 (6.8)	27.6 (6.8)
Wine	0.0 (0.0)	15.6 (5.2)	20.0 (6.4)	22.7 (6.0)	19.3 (7.9)	21.5 (3.7)	**25.2** (4.5)	**25.2** (4.5)	18.3 (7.2)
Average	19.4 (0.8)	34.2 (2.8)	40.3 (3.5)	40.5 (3.5)	**41.3** (4.4)	38.9 (5.2)	40.0 (3.1)	38.5 (3.2)	37.3 (3.6)

F-Measure for Logistic Regression (LR) [3] and SVM algorithms (averaged over 5 experiments).
"$I.R.$" I.R. based class costs; "G": Reproduction of [1] ; "C": $\mathbf{CONE}$ "T" thresholded predictions; "B": Bissection [2]

Perspectives

- Extend our study to more complex class of hypotheses (non linear hypotheses such as neural networks).
- Prove the convergence of our algorithm.
- Work on the notion of sub-optimality to improve the bound.
- Work on a generalization bound on the F-measure.

Acknowledgements

References

[1] S. P. Parambath, N. Usunier, and Y. Grandvalet, *Optimizing F-Measures by Cost-Sensitive Classification*, **NIPS** 2014.
[2] H. Narasimhan, H. Ramaswamy, A. Saha, and S. Agarwal, *Consistent Multiclass Algorithms for Complex Performance Measures*, **ICML** 2015.
[3] O. Koyejo, N. Natarajan, P. Ravikumar, and I. . Dhillon, *Consistent Binary Classification with Generalized Performance Metrics*, **NIPS** 2014.