UMR IRISA

Deep learning from imbalanced data

Elisa Fromont
(Thanks to Kevin Bascol, Rémi Emonet, Amaury Habrard,
Guillaume Metzler and Marc Sebban)

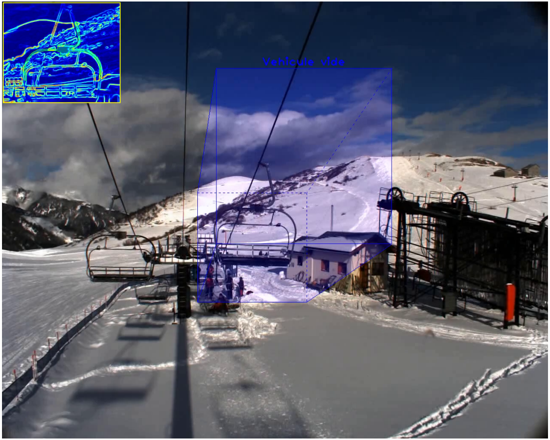
UNIVERSITÉ DE RENNES 1

6 September 2018
Conference - Deep Learning : from theory to applications
Labex Henri Lebesgue

A concrete application: anomaly detection in chairlift videos

Anomaly detection problem casts into a **binary image classification** problem (anomaly presence / absence in each track)

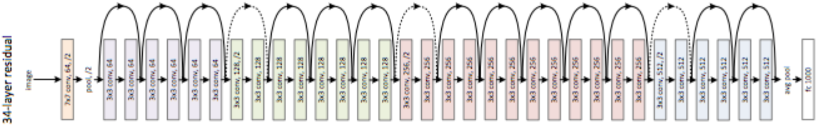
bluecime



Imbalanced Dataset (29 064 tracks)

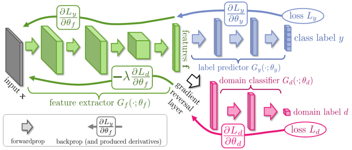
Outdoor scenes
Different chairlifts (15)
Different chairs, point of views
Different light, weather conditions
Neg = (53.9% **Empty**,38.7% **Safe**)
Pos = 7.4% **Unsafe**

1 - Find the right architecture (Resnet 50 + Data augmentation)



Deep residual learning for image recognition. He et al, CVPR 2016

2 - Few data on new domains) (Dom Adaptation + active learning)



Y.Ganin et al. Unsupervised domain adaptation by backpropagation. ICML 2015.

3 - Imbalanced data

$\rightarrow$ Very few positive examples must be as important as all negatives

Accuracy vs F-measure

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

In an imbalanced setting ($P \ll N$):

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 $\rightarrow$ difficult to optimize

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

Idea from Parambath et al. (NIPS 2014)

Optimizing F-Measures by Cost-Sensitive Classification, Parambath et al. NIPS 2014.

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

$\rightarrow$ Idea: find a link between the weighted error and the F-Measure.

Base Result (from Parambath et al .2014)

→ 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 $\boldsymbol{a'} \in \mathbb{R}_{+}^{d}$ and $h \in \mathcal{H}$ satisfying the following two conditions:

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

We have:

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

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

Geometric Interpretation

A weighting function:

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

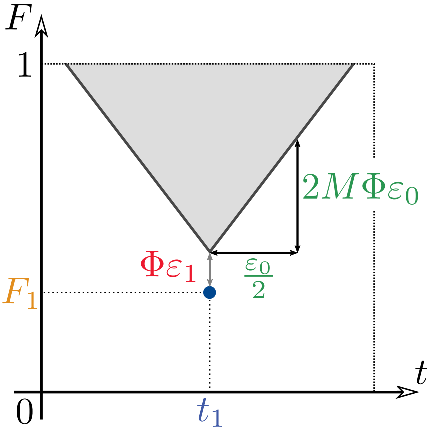
We can rewrite (using the Lipschitz property of a) the point

$\|\boldsymbol{a'} - \boldsymbol{a}^\star\|_2 \leq \varepsilon_0$, as

$\|a(t') - a(t^\star)\|_2 \leq 2\|t' - t^\star\|_2$ $(= \varepsilon_0)$,

and the bound in function of t:

$$\begin{aligned} F(\boldsymbol{e}(t^\star)) \leq\ & F(\boldsymbol{e}(t')) \\ & + 4\Phi M \|t' - t^\star\|_2 \\ & + \Phi\varepsilon_1. \end{aligned}$$



A Tighter Slope

$\rightarrow$ 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:

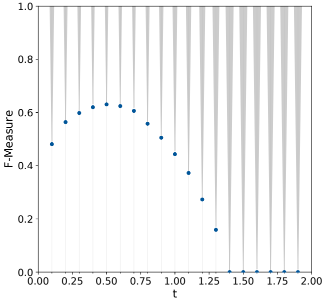
$$F(\boldsymbol{e}(t)) \leq F(\boldsymbol{e}(t_1)) + \Phi\sqrt{2}(\|\boldsymbol{e}\|_2 + M')\|t_1 - t\|_2 + \Phi\varepsilon_1.$$

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

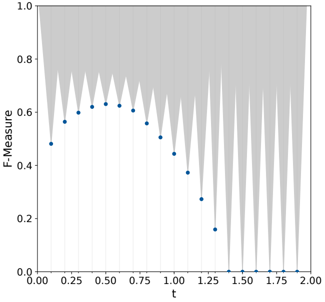
$$M' = \max_{\boldsymbol{e'} \in \mathcal{E}(\mathcal{H})} \|\boldsymbol{e'}\|_2, \quad s.t.\ F_\beta(\boldsymbol{e'}) > F_\beta(\boldsymbol{e}).$$

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

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

Search Space Pruning

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

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

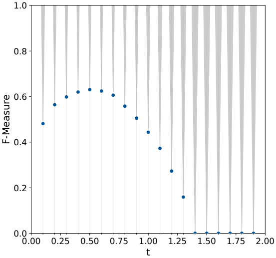
A bound on F-Measure

Let $t'' < t' < t$, and $\boldsymbol{e}(t), \boldsymbol{e}(t')$ and $\boldsymbol{e}(t'')$ the error profiles obtained with an optimal classifier trained with costs $a(t), a(t')$ and $a(t'')$ respectively. We have:

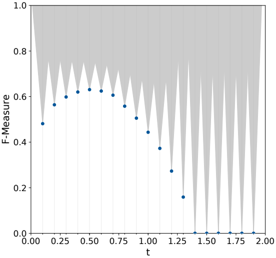
$$F_\beta(\boldsymbol{e}(t'')) \leq \frac{(1+\beta^2)P}{(1+\beta^2)P + e_2(t') - e_1(t')} \quad (1)$$

$$\text{and} \quad F_\beta(\boldsymbol{e}(t)) \leq \frac{(1+\beta^2)\left(P + e_2(t') - e_1(t')\right)}{(1+\beta^2)P + e_2(t') - e_1(t')}. \quad (2)$$

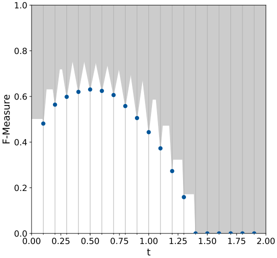
Unreachable region obtained with the bounds on points from a grid



Parambath et al.



CONE



CONE+ pruning

Presentation of **CONE**

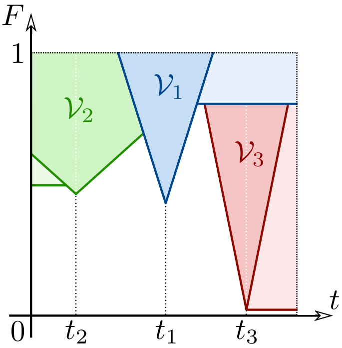


Illustration of **CONE** on the three first iterations.

$\boldsymbol{\nu}_1$: 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 $t \in [1.3,\ 2]$

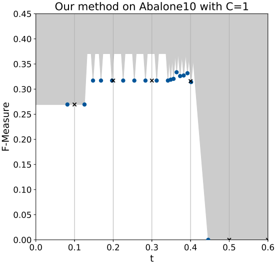
$\boldsymbol{\nu}_3$: Next cone with t in the middle of this interval: $t_3 = 1.65$

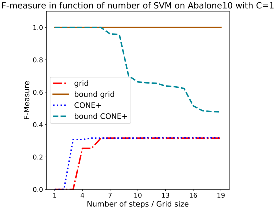
$\rightarrow$ Highest remaining $F = 0.7$ for $t \in [1.3,\ 1.35]$

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

Experimental Results

Abalone 10

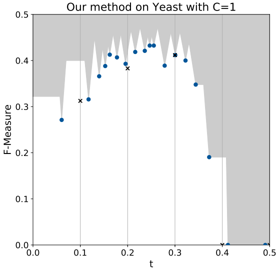


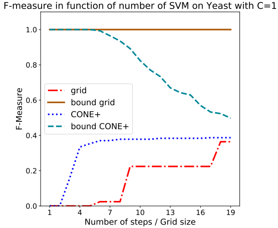


Examples of a run of **CONE** (blue points and gray area) compared to a simple grid search (black crosses)

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

Yeast





F-measure results on more and more imbalanced datasets

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

- ... 9 SVMs:

Dataset	SVM_G	SVM_C	SVM_C^r	SVM_C^l	SVM_C^{lr}
Adult	66.5 (0.1)	66.5 (0.0)	66.5 (0.1)	66.5 (0.1)	66.5 (0.1)
Abalone10	30.8 (1.2)	31.0 (1.1)	32.2 (1.3)	30.8 (1.2)	32.2 (1.0)
IJCNN'01	61.6 (0.5)	61.0 (0.6)	61.7 (0.6)	61.4 (0.8)	61.5 (0.7)
Abalone12	16.5 (2.6)	12.2 (7.0)	16.8 (4.6)	8.2 (7.3)	17.5 (4.5)
Yeast	36.8 (9.8)	34.8 (8.3)	38.9 (7.2)	33.7 (12.1)	37.8 (8.5)
Wine	18.4 (3.2)	11.3 (10.8)	16.0 (3.8)	14.5 (9.2)	16.9 (5.1)

- ... 4 SVMs:

Dataset	SVM_G	SVM_C	SVM_C^r	SVM_C^l	SVM_C^{lr}
Adult	66.5 (0.1)	66.5 (0.0)	66.5 (0.1)	66.5 (0.1)	66.5 (0.1)
Abalone10	30.8 (1.2)	12.2 (14.5)	30.9 (1.2)	21.2 (11.5)	30.8 (1.2)
IJCNN'01	59.8 (0.3)	61.0 (0.6)	61.4 (0.8)	61.4 (0.8)	61.6 (0.6)
Abalone12	0.0 (0.0)	0.0 (0.0)	15.7 (3.8)	2.8 (5.6)	15.7 (3.8)
Yeast	38.2 (11.7)	14.7 (12.0)	37.2 (7.4)	22.1 (16.4)	35.6 (8.8)
Wine	0.0 (0.0)	0.0 (0.0)	20.4 (7.6)	9.4 (11.7)	20.4 (7.6)

SVM_C: Reproduction of Parambath et al. algorithm

SVM_C: **CONE**

SVM_C^{lr}: **CONE** with left and right pruning

Conclusion & Future work

In this work,

- we derive a tighter bound on the F-measure,
- we propose an algorithm which prunes the search space of possible weights,
- we show empirically quick convergence results

But...,

- How can we apply this to neural networks? (what is ε_1 in this case?, training is expensive...)
- How can we have some generalization guarantees over F_β?

Questions?