Similarity Learning for Pointwise ROC Optimization

Robin Vogel[1,3], Stéphan Clémençon[1] and Aurélien Bellet[2]   [1] Télécom ParisTech, [2] Inria, [3] IDEMIA

Motivation

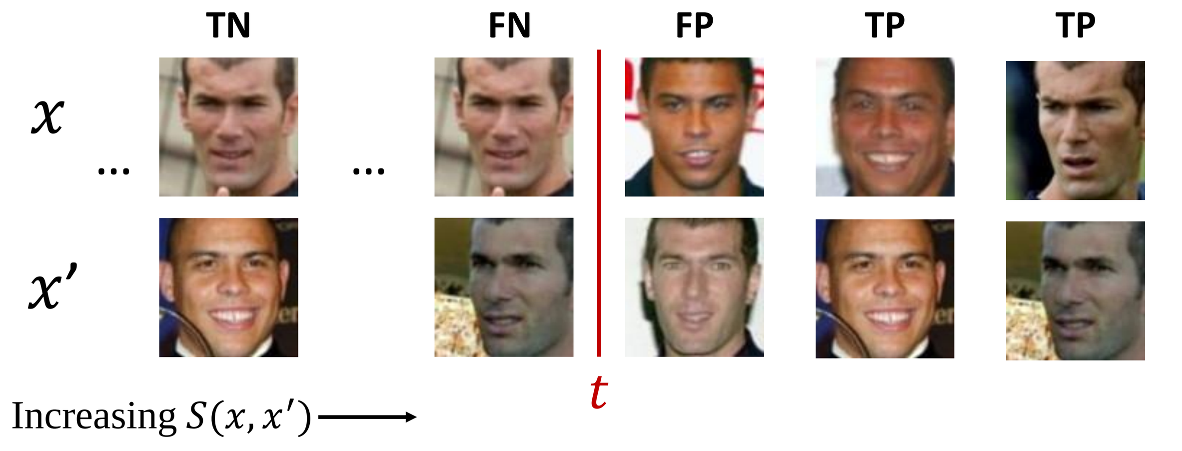
Biometric identification =
checks correspondance of two measurements (x, x').

Given a similarity S and a threshold t,

$$(x, x') \text{ is a match } \Leftrightarrow S(x, x') > t. \quad (1)$$

The ROC curve of S gives the true positive rate (TPR) given the false positive rate (FPR) associated to eq. (1).

Biometric systems are deployed to function at fixed FPR, see [1], hence we study *pointwise* ROC *optimization*.



Contributions

Proposition of an appropriate probabilistic framework for a novel perspective on similarity learning.

Statistical guarantees for the constrained optimization problem corresponding to the empirical version of our theoretical objective, i.e. pointwise ROC optimization in the context of pairwise ranking.

Faster rates under a weak low-noise assumption.

Empirical illustration of the faster rates through numerical simulations.

Study of sampling strategies for scalability issues that arise from the high number of negative pairs in practical settings.

References

[1] Anil K. Jain, Arun A. Ross, and Karthik Nandakumar. *Introduction to Biometrics*. Springer, 2011.

[2] Stephan Clémençon and Nicolas Vayatis. Overlaying classifiers: A practical approach to optimal scoring. *Constructive Approximation*, 32(3):619–648, 2010.

[3] Stephan Clémençon, Gàbor Lugosi, and Nicolas Vayatis. Ranking and Empirical Minimization of U-Statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

Preliminaries

Classification setting. Assume $(X, Y) \sim P$, with:

- $Y \in \{1, \dots, K\}$ the output label,
- $X \in \mathcal{X} \subset \mathbb{R}^d$ input random variable.

Similarity learning. Select a similarity function S s.t.

the larger $S(X, X')$ the higher $\mathbb{P}\{Y = Y' \mid X, X'\}$,

with $(X, Y) \perp (X', Y') \sim P$.

Optimal similarity rules $\mathcal{S}^*$ are increasing transforms of:

$$\eta(x, x') = \mathbb{P}\{Y = Y' \mid (X, X') = (x, x')\}.$$

Pointwise ROC optimization. Given target $\alpha \in (0, 1)$,

$$\max_{G \in \mathcal{G}_0} R^+(G) \quad \text{subject to} \quad R^-(G) \le \alpha, \quad (2)$$

where $\mathcal{G}_0$ class of fun, $R^+(G) = \mathbb{E}[G(X, X') \mid Y = Y']$ and $R^-(G) = \mathbb{E}[G(X, X') \mid Y \neq Y']$.

By Neyman-Pearson lemma:

$$G^*_\alpha := \mathbb{I}_{\mathcal{R}^*_\alpha} \text{ optimal solution of } \textbf{eq. (2)},$$

where $\mathcal{R}^*_\alpha = \{(x, x') \in \mathcal{X}^2 : \eta(x, x') \ge Q^*_\alpha\}$, with Q^*_α quantile of $\eta(X, X') \mid Y \neq Y'$ at level $1 - \alpha$.

Empirical problem. Using a training sample:

$$\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\},$$

composed of n i.i.d. copies of (X, Y), we form estimates of $R^+(G)$ and $R^-(G)$:

$$\hat{R}^+_n(G) = \frac{1}{n_+} \sum_{1 \le i < j \le n} G(X_i, X_j) \cdot \mathbb{I}\{Y_i = Y_j\},$$

$$\hat{R}^-_n(G) = \frac{1}{n_-} \sum_{1 \le i < j \le n} G(X_i, X_j) \cdot \mathbb{I}\{Y_i \neq Y_j\},$$

where $n_+ = \sum_{1 \le i < j \le n} \mathbb{I}\{Y_i = Y_j\} = n(n-1)/2 - n_-$.

One can then derive the empirical version of **eq. (2)**:

$$\max_{G \in \mathcal{G}_0} R^+_n(G) \quad \text{subject to} \quad R^-_n(G) \le \alpha + \Phi, \quad (3)$$

where $\Phi > 0$ is some tolerance parameter.

Generalization

Generalization guarantees. Our theorem describes the generalization capacities of the solution of **eq. (3)**, under some conditions on $\mathcal{G}_0$ and a suitable choice of Φ.

Theorem 1. *Suppose that:*

- *$\mathcal{G}_0$ is a VC-major class of VC-dimension V,*
- *$\forall G \in \mathcal{G}_0$, $\|G\|_\infty \le 1$,*
- *$\exists \kappa \in (0, 1)$ such that $\kappa \le \mathbb{P}\{Y = Y'\} \le 1 - \kappa$,*

For all $\delta \in (0, 1)$ and $n > 1$, :

- *set $\Phi_{n,\delta} = C_{V,\delta,\kappa} \cdot n^{-1/2}$,*
  where $C_{V,\delta,\kappa}$ is known and depends on V, δ, κ,
- *Let $\hat{G}_n$ solution of* **eq. (3)** *with $\Phi = \Phi_{n,\delta/2}$,*

We have w.p. $\ge 1 - \delta$, $\forall n \ge 1 + 4\kappa^{-2}\log(3/\delta)$,

$$R^+(\hat{G}_n) \ge \sup_{G \in \mathcal{G}_0 :\, R^-(G) \le \alpha} R^+(G) - \Phi_{n,\delta/2},$$

$$\text{and} \quad R^-(\hat{G}_n) \le \alpha + \Phi_{n,\delta/2}.$$

Fast Rates

In some situations, rates faster than $O(1/\sqrt{n})$ can be achieved by solutions of **eq. (3)**. These rates hold when the following noise assumption is verified:

Noise assumption (NA). $\exists c \in \mathbb{R}^*_+, a \in [0, 1]$ *s.t.,*

$$\mathbb{E}_{X'}\left[|\eta(X, X') - Q^*_\alpha|^{-a}\right] \le c \quad a.s.$$

Our next theorem establishes fast rate bounds under the **(NA)** condition on the data distribution. It relies on a variant of the Bernstein inequality for U-statistics.

Theorem 2. *Suppose that:*

- *the assumptions of* **theorem 1** *are satisfied,*
- *NA holds true,*
- *$G^*_\alpha \in \mathcal{G}_0$,*

Fix $\delta > 0$, there exists C', that depends of δ, κ, Q^*_α, a, c and V such that, w.p. $\ge 1 - \delta$,

$$R^+(\hat{G}_n) \ge R^+(G^*_\alpha) - C' \cdot n^{-(2+a)/4},$$

$$\text{and} \quad R^-(\hat{G}_n) \le \alpha + \Phi_{n,\delta/2}.$$

Scalability

When n large and K large, calculating $R^-_n(G)$ is computationally costly. A sensible approach is to drastically subsample the negative pairs, while keeping all positive pairs.

We studied the equivalent of **eq. (3)** when replacing $\hat{R}^-_n(G)$ by the following approximation:

$$\bar{R}^-_B(G) := \frac{1}{B} \sum_{(i,j) \in \mathcal{P}_B} G(X_i, X_j),$$

where $\mathcal{P}_B$ is a set of cardinality B built by sampling with replacement in the set of negative training pairs Λ_P, with:

$$\Lambda_P = \{(i, j) \mid i, j \in \{1, \dots, n\}; Y_i \neq Y_j\}.$$

We show that the results of **theorem 1** still hold, with a different Φ of the order $O(\sqrt{\log(n)/B})$.

It implies that it is sufficient to sample $B = O(n)$ pairs to get a learning rate of order $O(\sqrt{\log(n)/n})$.

Experiment on Fast Rates

We illustrate the results presented in **theorem 2**.

How ?
Introduce distributions that satisfy **(NA)** with different a's, and show the difference in the generalization speed.

Defining $K = 2$, $X \sim \mathcal{U}[0, 1]$, $\mathbb{P}\{Y = 1\} = 1/2$ and the density μ_1 of $X \mid Y = 1$ fully caracterizes P. Hence we define a family of μ_1's parameterized by a.

For some a, the 90-quantile of $\log(R^+(G^*_\alpha) - R^+(\hat{G}_n))$ for different n's is fitted to $C_a \times \log(n) + D_a$ to get the empirical generalization speed C_a.