On Tree-based Methods for Similarity Learning

Stephan Clémençon[1] **Robin Vogel**[1,2]

[1] Telecom Paris, [2] IDEMIA

LOD 2019 - 11/09/2019

Outline

Introduction

TreeRank for bipartite ranking

Similarity TreeRank

Biometric verification (1/2)

$S(X, X') > t$

A biometric system uses:

- Two **measurements** X and X',
- A **similarity** S that quantifies the likeness of (X, X'),
- A **threshold** t that separates positive and negative pairs.

Aim: $S(X, X') > t$ is a good indicator of $Z = +1$ with:

$$Z = \begin{cases} +1 & \text{if } (X, X') \text{ from the same person,} \\ -1 & \text{otherwise.} \end{cases}$$

Two types of errors:

$$\begin{aligned} \mathrm{TPR}_S(t) &:= \mathbb{P}\{S(X, X') > t \mid Z = +1\}, \\ \mathrm{FPR}_S(t) &:= \mathbb{P}\{S(X, X') > t \mid Z = -1\}. \end{aligned}$$

The set $\{(FPR_S(t), TPR_S(t)) \mid t \in \mathbb{R}\}$ is known as the **ROC curve**.

Biometric verification (2/2)

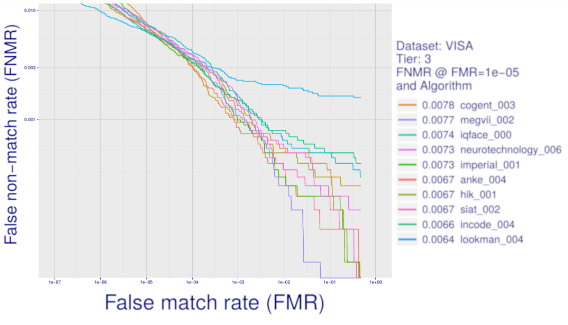


Figure: Extract of the NIST Face Recognition Vendor Test (FRVT) report.

Several criterions measure ROC accuracy:

- Area Under the ROC Curve (**AUC**),
- Pointwise ROC optimization (**pROC**) see [Vogel et al., 2018],
- Local AUC (**LocAUC**): see [Clémençon and Vayatis, 2007].

Remarks:

· **AUC** does not focuses on best instances,

· **pROC** and **LocAUC** can have inadapted optimums.

Our contribution

The posterior probability $\eta(x, x') = \mathbb{P}\{Z = +1 | (X, X') = (x, x')\}$, is the **optimal similarity** in a ranking and ROC sense.

Contributions:

- Procedure for building a **similarity that approximates η.**
- Guarantees in **sup norm** in the **ROC space**.
- Implementations in specific cases.

Plan:

1. Present **TreeRank for bipartite ranking**, see [Clemencon and Vayatis, 2009].
2. Adapt it to similarity learning.
3. Draw on **U-statistic theory** to prove new results.

Outline

Bipartite ranking

Standard **binary classification** framework:

- let a random variable $(X, Y) \in \mathcal{X} \times \{-1, +1\}$,
- n i.i.d. copies $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^{n}$ of (X, Y),
- the distribution of (X, Y) is summarized by (μ, η) where:

$$\forall x \in \mathcal{X},\ \eta(x) = \mathbb{P}\{Y = +1 | X = x\},\ \mu(C) = \mathbb{P}\{X \in C\}.$$

- or by (p, α, β), where $p = \mathbb{P}\{Y = +1\} = 1 - q$,

$$\alpha(C) = \mathbb{P}\{X \in C | Y = -1\} \text{ and } \beta(C) = \mathbb{P}\{X \in C | Y = +1\},$$

  α the **false positive rate (FPR)**, β the **true positive rate (TPR)**.

Objective: Rank items in $\mathcal{X}$ by decreasing η using $\mathcal{D}_n$.

The ranking is derived from **a scorer** $s : \mathcal{X} \to \mathbb{R}$ that ranks $\mathcal{X}$ with $\mathbb{R}$.

Optimal scorers $\mathcal{S}^*$ are **increasing transforms** $T \circ \eta$ of η.
A scorer s^* is optimal i.f.f. $\exists w \in \mathcal{L}_1, w \geq 0$ and V cont. r.v. in $(0, 1)$:

$$\forall x \in \mathcal{X}, \quad s^*(x) = \inf_{z \in \mathcal{X}} s^*(z) + \mathbb{E}\left[w(V) \cdot \mathbb{I}\{\eta(x) > V\}\right].$$

An approach to bipartite ranking (1/2)

Idea: Using **estimated super-level sets** of η: $\{x \in \mathcal{X} | \eta(x) > t\}_{t \in \mathbb{R}}$, build an **accurate scorer**.

Optimal scorers write as: $s^*(x) = C + \mathbb{E}\left[w(V) \cdot \mathbb{I}\{\eta(x) > V\}\right]$, which can be **approximated by a piecewise constant function**,

$$s_N(x) = \sum_{j=1}^{N} \mathbb{I}\{x \in R_j\},$$

with $\{R_j\}_{1=1}^{N}$ **increasing** ($R_j \subset R_{j+1}$) family of sets.

The ROC curve of s_N is the broken line connecting the dots:

$$\{(\alpha(R_j), \beta(R_j))\}_{0 \leq j \leq N}. \text{ with } R_0 = \emptyset. \tag{1}$$

Hence, if the R_j's are the level sets of η,
eq. (1) could be a **piecewise linear** approximation of the ROC curve.

An approach to bipartite ranking (2/2)

From the **Neymann-Pearson fundamental lemma**, the optimal solution of pROC at level α, i.e.

$$\max_{C} \beta(C) \text{ s.t. } \alpha(C) \leq \alpha, \tag{2}$$

is $\{x \in \mathcal{X} | \eta(x) > \gamma\}$ where γ is the $(1-\alpha)$-quantile of $\eta(X)|Y = -1$.

[Clémençon and Vayatis, 2009] solves eq. (2) to build good scorers.

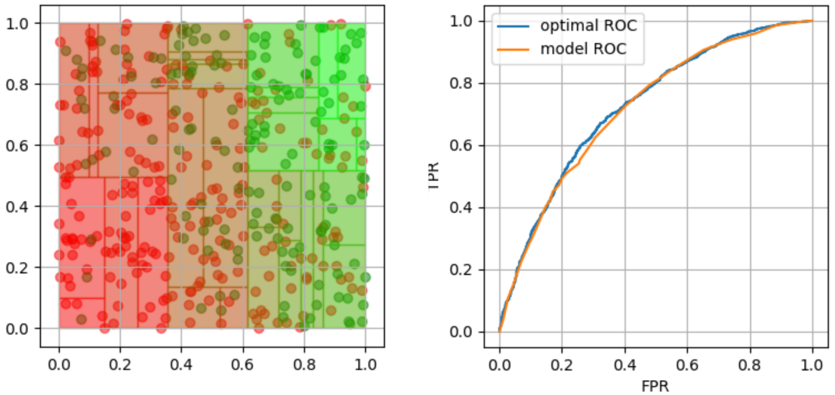
The **weighted classif loss**: $L_c(C) = cp \cdot \beta(C) - (1-c)q \cdot \alpha(C)$, has for optimal solution $C = \{x \in \mathcal{X} | \eta(x) > 1 - c\}$.

TreeRank [Clemencon and Vayatis, 2009] exploits that idea.

TreeRank splits $\mathcal{X}$ **recursively** to retrieve the super-level sets of η, with **weighted classif.** optimized on a **family** $\mathcal{C}$ **of subsets of** $\mathcal{X}$**.**

Visual example of TreeRank

For $\mathcal{C}$ being **coordinate splits**, $\mathcal{X} = [0,1]^2$ and $\eta(x) = (4x_1 + 2x_2)/7$:



Remark: Elements of $\mathcal{C}$ are very different from super-level sets of η.

TreeRank algorithm

Input. Maximal depth $D \geq 1$ of the tree, $\mathcal{C}$, $\mathcal{D}_n$.

1. (INIT.) Set $\mathcal{C}_0 = \mathcal{X}$, $\alpha_{d,0} = \beta_{d,0} = 0$ and $\alpha_{d,2^d} = \beta_{d,2^d} = 1$ for all $d \geq 0$.
2. (ITERATIONS.) For $d = 0, \ldots, D-1$ and $k = 0, \ldots, 2^d - 1$:

   2.1 (OPTIMIZATION STEP.) Set the entropic measure:

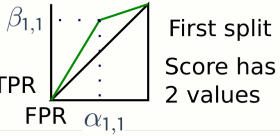
   $$\widehat{\Lambda}_{d,k+1}(C) = (\alpha_{d,k+1} - \alpha_{d,k}) \cdot \hat{\beta}(C) - (\beta_{d,k+1} - \beta_{d,k})\hat{\alpha}(C).$$

   Find $\mathcal{C}_{d+1,2k} \in \mathcal{C}$ subset of $\mathcal{C}_{d,k}$ that maximizes $\widehat{\Lambda}_{d,k+1}$.

   2.2 (UPDATE.) Set $\alpha_{d+1,2k+1}$, $\beta_{d+1,2k+1}$, $\alpha_{d+1,2k+2}$ and $\beta_{d+1,2k+2}$.
3. (OUTPUT.) After D iterations, get the piecewise constant score function:

$$s_D(x) = \sum_{k=0}^{2^D-1} (2^D - k)\mathbb{I}\{x \in \mathcal{C}_{D,k}\},$$

- For **pruning**: [Clemencon and Vayatis, 2009].
- For **bagging**: [Clémençon et al., 2013].



Guarantees

Assumptions: Let $\prec$ denote **absolute continuity**, [1]

1. α and β are **equivalent** (i.e. $\alpha \prec \beta$ and $\beta \prec \alpha$), $\frac{d\beta}{d\alpha}$ is bounded.
2. $\eta(X) \prec \lambda$ with λ the Lebesgue measure.
3. $\mathcal{C}$ is of VC-dimension V,
4. contains all level sets of η and is intersection-stable ($C \cap C' \in \mathcal{C}$),

Theorem: Under those, given a tree of depth $D = D_n \sim \log(\sqrt{n})$, $\forall \delta > 0, \exists \lambda > 0$ such that, with probability $\geq 1 - \delta, \forall n \in \mathbb{N}$,

$$\left\| \mathrm{ROC}_{s_{D_n}} - \mathrm{ROC}_{s^*} \right\|_\infty \leq \exp(-\lambda \sqrt{\log(n)}).$$

λ depends on V, δ and universal constants.

---

[1] $\mu \prec \nu$ i.f.f. $\exists h : \mathcal{X} \to \mathbb{R}^+, \mu = h \cdot \nu$.

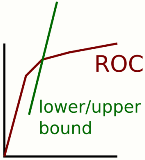
Sketch of proof ([Clemencon and Vayatis, 2009])

Step 1: Bound the $\|\cdot\|_\infty$

With $\beta^*_{D,k}, \alpha^*_{D,k}$ params for the *adaptative broken line* est. of ROC_{s^*}, when $t \in [\alpha^*_{D,k_0}, \alpha_{D,k_0+1}] \cap [\alpha(R_{D,k-1}), \alpha(R_{D,k})]$,

$$\mathrm{ROC}_{s^*}(t) \leq \beta^*_{D,k_0} + \mathrm{ROC}'_{s^*}(0) \times (t - \alpha^*_{D,k_0}),$$

$$\mathrm{ROC}_{s_D}(t) \geq \beta(R_{D,k}) - \mathrm{ROC}'_{s^*}(0) \times (t - \alpha(R_{D,k})).$$



It implies the following bound on $\|\mathrm{ROC}_{s^*} - \mathrm{ROC}_{s_D}\|_\infty$:

$$\max_{1\leq k \leq 2^D-1} \beta^*_{D,k} - \beta(R_{D,k-1}) + \mathrm{ROC}'_{s^*}(0) \left[\alpha^*_{D,k} - \alpha(R_{D,k-1})\right]. \quad (3)$$

Step 2: Prove by recurrence a bound of eq. (3)

Under assumptions, $\exists K$ s.t. $\forall \delta > 0$, with probability $1-\delta$, $\forall d, k$:

$$|\alpha^*_{d,k} - \alpha(R_{d,k-1})| + |\beta^*_{d,k} - \beta(R_{d,k-1})| \leq K^d B(d+1, n, \delta),$$

where $B(d+1,n,\delta) = O\left(\frac{V^{1/2d} + \log(1/\delta)^{1/2d}}{n^{1/2d}}\right)$.

Which is proven using **standard VC inequalities**.

Outline

Similarity ranking

We chose the standard **classification** framework:

- let a random variable $(X, Y) \in \mathcal{X} \times \{1, \ldots, K\}$,
- the distribution of (X, Y) is summarized by $(\mu, (\eta_1, \ldots, \eta_K))$:

$$\forall x \in X, k \in \{1, \ldots, K\}, \quad \eta_k(x) = \mathbb{P}\{Y = k | X = x\}.$$

- the optimal similarity is $\eta(x, x') = \sum_{k=1}^{K} \eta_k(x)\eta_k(x')$, i.e. the probability to be in the same class.

Objective: Rank pairs in $\mathcal{X} \times \mathcal{X}$ by decreasing η using $\mathcal{D}_n$.

The ranking is derived from **a similarity** $s : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

Given two i.i.d. pairs (X, Y) and (X', Y'), set $Z = 2\mathbb{I}\{Y = Y'\} - 1$, one can form $n(n-1)/2$ obs. of the form $((X, X'), Z)$ from $\mathcal{D}_n$.

Idea: Run **TreeRank on non-i.i.d. data** of the form $((X, X'), Z)$.

TreeRank for similarity ranking

TreeRank on data of the form $((X, X'), Z)$ gives a similarity s.

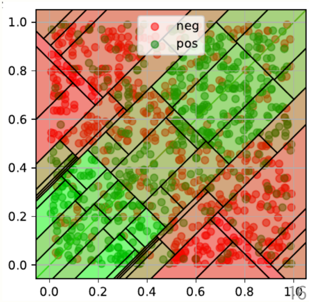
Similarities satisfies **more constraints** than scorers.

Symmetricity: For s to be symmetric, it suffices that $\mathcal{C}$ is symmetric.
$\rightarrow$ Use symmetric proposal regions,
$\rightarrow$ Learn on data of the form $((X + X', |X - X'|), Z)$.

Lemma: Let $s : \mathcal{X}^2 \rightarrow \mathbb{R}$,
s is symmetric $\Leftrightarrow \exists s_0 : \mathcal{X}^2 \rightarrow \mathbb{R},\ s(x, x') = s_0(x + x', |x - x'|)$.

Identity:
One expects $s(x, x) \geq s(x, z)$ for all $x, z \in \mathcal{X}$.
It is not satisfied by:

$$\eta(x, x') = \sum_{k=1}^{K} \eta_k(x)\eta_k(x').$$



Extension of the results

Same type guarantees hold for similarity TreeRank.

Estimates of α, β are ratios of *U*-**statistics** (means of pairs), e.g.:

$$\hat{\alpha}(C) = \frac{1}{n_-} \sum_{i<j} \mathbb{I}\{Z_{i,j} = -1, (X_i, X_j) \in C\},$$

and **standard VC inequalities do not hold.**

Using the fact that a simple U-statistic $U_n(h)$ can be rewritten as:

$$U_n(h) := \frac{2}{n(n-1)} \sum_{i<j} h(X_i, X_j) = \frac{1}{n!} \sum_{\sigma \in \mathfrak{S}_n} \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} h(X_{\sigma(i)}, X_{\lfloor n/2 \rfloor + i}),$$

which is the **first Hoeffding decomposition**, Jensen's inequality implies new VC inequalities, see [Vogel et al., 2018].

Bagging TreeRank models learned on **incomplete U-stats** works, see [Clémençon et al., 2016].

Conclusion

In a nutshell:

TreeRank describes a **general approach to the ranking problem**.

Performance of **similarity learning** algorithms is evaluated by **ranking criterions** in important applications.

TreeRank can be adapted for **similarity learning**, and theoretical results extended **using results on U-statistics**, see [Clémençon et al., 2016].

TreeRank's competitiveness depends on **the expressivity of $\mathcal{C}$**.

Future work:

Explore the idea of **using neural networks** to represent $\mathcal{C}$.

Merci !

References

Clémençon, S., Colin, I., and Bellet, A. (2016).
Scaling-up Empirical Risk Minimization: Optimization of Incomplete *U*-statistics.
Journal of Machine Learning Research, 17(76):1–36.

Clémençon, S. and Vayatis, N. (2007).
Ranking the best instances.
Journal of Machine Learning Research, 8:2671–2699.

Clémençon, S. and Vayatis, N. (2009).
Overlaying classifiers: a practical approach for optimal ranking.
In *Constructive Approximation*, volume 32, pages 313–320.

Clémençon, S., Depecker, M., and Vayatis, N. (2013).
Ranking Forests.
Journal of Machine Learning Research, 14:39–73.

Clemencon, S. and Vayatis, N. (2009).
Tree-based ranking methods.
IEEE Transactions on Information Theory, 55(9):4316–4336.

Vogel, R., Bellet, A., and Clémençon, S. (2018).
A probabilistic theory of supervised similarity learning for pointwise ROC curve optimization.
In *ICML 2018*.