Algorithmic Robustness for Semi-Supervised (ϵ, γ, τ)-Good Metric Learning

Maria-Irina Nicolae[1,2] Marc Sebban[1] Amaury Habrard[1] Éric Gaussier[2] Massih-Reza Amini[2]

[1]Université Jean Monnet, Laboratoire Hubert Curien, France [2]Université Grenoble Alpes, CNRS-LIG/AMA, France

Abstract

The notion of metric plays a key role in machine learning problems such as classification, clustering or ranking. However, it is worth noting that there is a severe lack of theoretical guarantees that can be expected on the generalization capacity of the classifier associated to a given metric. The theoretical framework of (ϵ, γ, τ)-**good similarity functions** [1] has been one of the first attempts to draw a link between the properties of a similarity function and those of a linear classifier making use of it. We extend this theory by providing a new **generalization bound** for the associated classifier based on the **algorithmic robustness** framework.

Problem setting

- Labeled examples $(\mathbf{x}, l(\mathbf{x}))$ drawn from some unknown distribution P over $\mathcal{X} \times \{-1, 1\}$, where $\mathcal{X} \subseteq \mathbb{R}^d$;
- Unlabeled examples $\mathbf{x}$ drawn from P over $\mathcal{X}$;
- Generic similarity function $K_{\mathbf{A}} : \mathcal{X} \times \mathcal{X} \to [-1, 1]$ over $\mathcal{X}$, possibly parameterized by a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$;
- Learning a large margin global separator $\boldsymbol{\alpha}$;
- Providing theoretical guarantees depending on $K_{\mathbf{A}}$.

(ϵ, γ, τ)-Good Similarity Functions [1]

Definition 1. $K_{\mathbf{A}}$ is a (ϵ, γ, τ)-**good similarity function** in hinge loss for a learning problem P if there exists a random indicator function $R(\mathbf{x})$ defining a probabilistic set of "reasonable points" such that the following conditions hold:

1. $\mathbb{E}_{(\mathbf{x}, l(\mathbf{x})) \sim P}\left[[1 - l(x)g(\mathbf{x})/\gamma]_+\right] \leq \epsilon$, where $g(\mathbf{x}) = \mathbb{E}_{(\mathbf{x}', l(\mathbf{x}'), R(\mathbf{x}'))}\left[l(\mathbf{x}')K_{\mathbf{A}}(\mathbf{x}, \mathbf{x}')|R(\mathbf{x}')\right]$.
2. $\Pr_{\mathbf{x}'}(R(\mathbf{x}')) \geq \tau$.

Theorem 2. *Using similarity scores to reasonable points as features, there exists a linear separator $\boldsymbol{\alpha}$ that has error ϵ at margin γ.*

Formulation

$$\min_{\boldsymbol{\alpha}} \frac{1}{d_l} \sum_{i=1}^{d_l} \ell(\mathbf{A}, \boldsymbol{\alpha}, \mathbf{z}_i) \quad \text{s.t.} \quad \sum_{j=1}^{d_u} |\alpha_j| \leq 1/\gamma, \qquad (1)$$

where $\ell(\mathbf{A}, \boldsymbol{\alpha}, (\mathbf{x}_i, l(\mathbf{x}_i))) = \left[1 - \sum_{j=1}^{d_u} \alpha_j l(\mathbf{x}_i) K_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j)\right]_+$ is the instantaneous loss estimated at point $(\mathbf{x}_i, l(\mathbf{x}_i))$.

Prediction rule

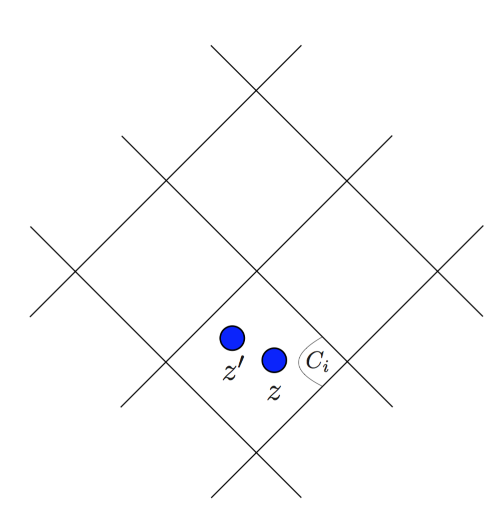
$$l(\mathbf{x}) = \operatorname{sgn} \sum_{j=1}^{d_u} \alpha_j K_{\mathbf{A}}(\mathbf{x}, \mathbf{x}_j)$$

- d_l: # of training examples
- d_u: # of unlabeled reasonable points

Algorithmic Robustness

Definition 3.[6] An algorithm is **robust** if for any example $\mathbf{z}'$ falling in the same subset as a training example $\mathbf{z}$, the gap between the losses associated with $\mathbf{z}$ and $\mathbf{z}'$ is bounded.

Theorem 4. *Given a partition of $\mathcal{Z}$ into M subsets $\{C_i\}$ and $K_{\mathbf{A}}(\mathbf{x}, \mathbf{x}')$, a k-lipschitz similarity function, Problem (1) is $(M, \frac{1}{\gamma}k\rho)$-robust with $\rho = \sup_{\mathbf{x}, \mathbf{x}' \in C_i} ||\mathbf{x} - \mathbf{x}'||$.*



k-lipschitz similarity functions

- $K^1_{\mathbf{A}}(\mathbf{x}, \mathbf{x}') = 1 - (\mathbf{x} - \mathbf{x}')^T \mathbf{A} (\mathbf{x} - \mathbf{x}')$, $k = 4||\mathbf{A}||_2$
- $K^2_{\mathbf{A}}(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{A} \mathbf{x}'$, $k = ||\mathbf{A}||_2$
- $K^3_{\mathbf{A}}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{(\mathbf{x}-\mathbf{x}')^T \mathbf{A} (\mathbf{x}-\mathbf{x}')}{2\sigma^2}\right)$, $k = \frac{2||\mathbf{A}||_2}{\sigma^2}\left(\exp\left(\frac{1}{2\sigma^2}\right) - \exp\left(\frac{-1}{2\sigma^2}\right)\right)$.

Learning Guarantees

Theorem 5. *For any $\delta > 0$ with probability at least $1 - \delta$, we have:*

$$\underbrace{|\mathcal{R}^\ell}_{\text{true risk}} - \underbrace{\widehat{\mathcal{R}}^\ell|}_{\text{empirical risk}} \leq \frac{1}{\gamma} \overbrace{k}^{\text{Lipschitz constant of } K_{\mathbf{A}}} \underbrace{\rho}_{\text{part size}} + \left(1 + \frac{1}{\gamma}\right) \sqrt{\frac{2\overbrace{M}^{\#\text{ of parts in partition}} \ln 2 + 2\ln(1/\delta)}{d_l}}.$$

Application to Joint Similarity Learning

Formulation of JSL

$$\min_{\boldsymbol{\alpha}, \mathbf{A}} \frac{1}{d_l} \sum_{i=1}^{d_l} \left[1 - \sum_{j=1}^{d_u} \alpha_j l(\mathbf{x}_i) K_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j)\right]_+ \qquad \text{s.t.} \sum_{j=1}^{d_u} |\alpha_j| \leq 1/\gamma$$

$$\mathbf{A} \text{ diagonal}, |A_{kk}| \leq 1, \quad 1 \leq k \leq d.$$

Table 1: Average accuracy of JSL (%) with CI at 95%, 5 labeled points per class, all points used as unlabeled.

Sim.	Balance	Ionosphere	Iris	Liver	Pima	Sonar	Wine
$K^1_{\mathbf{A}}$	85.7±3.5	88.5±2.6	**74.5**±4.4	63.9±5.3	71.1±3.8	**72.3**±4.1	**87.7**±5.0
$K^2_{\mathbf{A}}$	**87.1**±2.5	**91.0**±2.0	71.4±5.9	**69.2**±3.2	**72.9**±3.9	71.9±4.2	84.2±6.9
$K^3_{\mathbf{A}}$	81.1±8.5	86.2±2.8	68.2±8.5	58.6±6.3	71.1±4.3	63.9±10.0	83.5±6.2

Table 2: Average accuracy (%) over all datasets with CI at 95%.

Method	5 pts./cl.	10 pts./cl.	20 pts./cl.
3NN	64.6±4.6	68.5±5.4	70.4±5.0
LMNN-diag [5]	65.1±5.5	68.2±5.6	71.5±5.2
LMNN [5]	69.4±5.9	70.9±5.3	73.2±5.2
ITML [3]	75.8±4.2	76.5±4.5	76.3±4.8
SVM	76.4±4.9	76.2±7.0	77.7±6.4
BBS [1]	77.2±7.3	77.0±6.2	77.3±6.3
SLLC [2]	70.5±7.2	75.9±4.5	75.8±4.8
LRML [4]	74.7±6.2	75.3±5.9	75.8±5.2
JSL	**78.4**±2.3	**78.7**±1.9	**78.3**±1.6

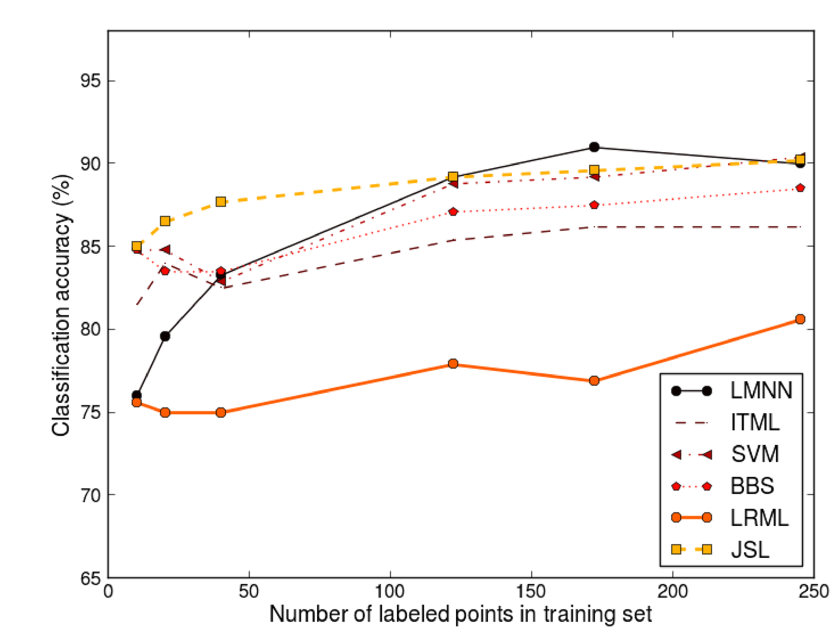


Figure 1: Ionosphere with 15 unlabeled points.

Conclusion

- New generalization bound for the (ϵ, γ, τ)-good framework;
- Generic form of similarity function;
- Experiments for learning the similarity with guarantees.

Acknowledgments Funding for this project was provided by a grant from Région Rhône-Alpes.

References

[1] M.-F. Balcan, A. Blum, and N. Srebro. Improved guarantees for learning via similarity functions. In R. A. Servedio and T. Zhang, editors, *COLT*, pages 287–298. Omnipress, 2008.

[2] A. Bellet, A. Habrard, and M. Sebban. Similarity learning for provably accurate sparse linear classification. In *ICML 2012*, pages 1871–1878, 2012.

[3] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216, New York, NY, USA, 2007. ACM.

[4] S. C. H. Hoi, W. Liu, and S.-F. Chang. Semi-supervised distance metric learning for collaborative image retrieval. In *CVPR*, 2008.

[5] K. Weinberger and L. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10:207–244, 2009.

[6] H. Xu and S. Mannor. Robustness and generalization. *Machine Learning*, 86(3):391–423, 2012.