Joint Semi-Supervised Similarity Learning for Linear Classification

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

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

Metric Learning

- Aims at optimizing parameterized distances/similarities.
- Leads to transformations of the input space before learning the classifier.
- Takes constraints from data.

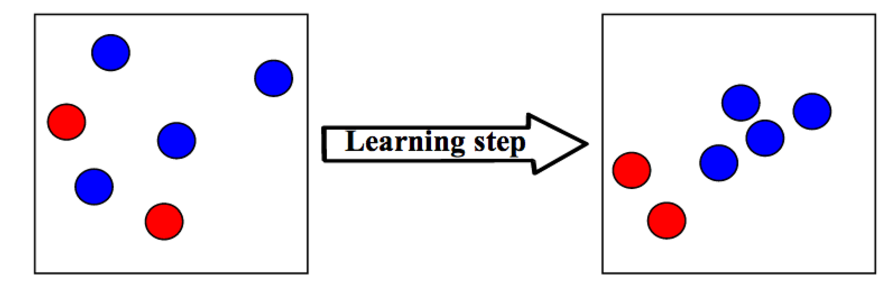
Mahalanobis Distance

Finds $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ positive semi definite (PSD) parameterizing $d_{\boldsymbol{A}}$, s.t. it best satisfies the constraints.

$$d_{\boldsymbol{A}}(\mathbf{x},\mathbf{x}') = \sqrt{(\mathbf{x}-\mathbf{x}')^T \boldsymbol{A}(\mathbf{x}-\mathbf{x}')}$$

Limitations

- Satisfying $\boldsymbol{A}$ PSD is computationally expensive.
- No generalization guarantees are provided.



(ϵ,γ,τ)-Good Similarity Functions [1]

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

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

Theorem 2. Using similarity scores to landmarks as features, there exists a linear separator α that has error ϵ at margin γ.

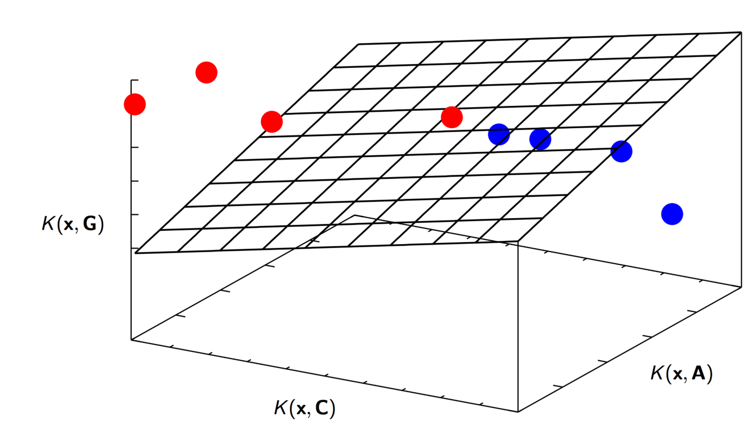
- d_l: # of training examples
- d_u: # of unlabeled landmarks

Formulation

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

Prediction rule

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



Joint Similarity and Classifier Learning

Goal Jointly optimize the empirical goodness of α and $K_{\mathbf{A}}$ from sample $\mathcal{S}$.

- $\mathbf{A}$ is not constrained to be PSD.
- Semi-supervised setting, averaged constraints.
- Solved by alternating optimization steps over α and $\mathbf{A}$.

Formulation of JSL

$$\min_{\alpha,\mathbf{A}} \sum_{i=1}^{d_l}\left[1-\sum_{j=1}^{d_u}\alpha_j y_i K_{\mathbf{A}}(\mathbf{x}_i,\mathbf{x}_j)\right]_+ + \lambda\|\mathbf{A}-\mathbf{R}\|$$

$$\text{s.t.} \sum_{j=1}^{d_u}|\alpha_j| \le 1/\gamma \text{ and } \mathbf{A} \text{ diagonal, } |A_{kk}| \le 1, \quad 1 \le k \le d$$

Regularizer $\|\mathbf{A}-\mathbf{R}\|$

- L_1 or L_2 norm
- Value of $\mathbf{R} \in \mathbb{R}^{d\times d}$
  - Identity matrix
  - Empirical estimate of Kullback-Leibler divergence

(β, c)-Admissibility

Definition 3. $K_{\mathbf{A}}$ is (β,c)-admissible if, for any matrix norm $\|\cdot\|$, there exist $\beta, c \in \mathbb{R}$ s.t. $\forall \mathbf{x},\mathbf{x}'$, $|K_{\mathbf{A}}(\mathbf{x},\mathbf{x}')| \le \beta + c\cdot\|\mathbf{x}'\mathbf{x}^T\|\cdot\|\mathbf{A}\|$.

- $K^1_{\mathbf{A}}(\mathbf{x},\mathbf{x}') = \mathbf{x}^T\mathbf{A}\mathbf{x}'$ is $(0,1)$-admissible;
- $K^2_{\mathbf{A}}(\mathbf{x},\mathbf{x}') = 1-(\mathbf{x}-\mathbf{x}')^T\mathbf{A}(\mathbf{x}-\mathbf{x}')$ is $(1,4)$-admissible.

Rademacher Complexity

Definition 4. $\mathcal{R}_n(\mathcal{F}) := \mathbb{E}_{\mathcal{S}}\mathbb{E}_{\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)\right], \forall n$

- $\mathcal{F}$ class of uniformly bounded functions
- $\{\sigma_i : i \in \{1,\dots,n\}\}$ independent Rademacher random variables, $\Pr(\sigma_i = 1) = \Pr(\sigma_i = -1) = \frac{1}{2}$

Learning Guarantees for JSL

Theorem 5. Let $(\mathbf{A}_{\mathcal{S}},\alpha_{\mathcal{S}})$ be the solution to JSL and $K_{\mathbf{A}_{\mathcal{S}}}$ a (β,c)-admissible similarity function. Then, for any $0<\delta<1$, with probability at least $1-\delta$, the following holds:

true risk — (β,c)-admissibility of $K_{\mathbf{A}}$ — $X_* = \sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}}\|\mathbf{x}'\mathbf{x}^T\|_*$

$$\mathcal{E}(\mathbf{A}_{\mathcal{S}},\alpha_{\mathcal{S}}) - \mathcal{E}_{\mathcal{S}}(\mathbf{A}_{\mathcal{S}},\alpha_{\mathcal{S}}) \le 4\mathcal{R}_{d_l}\left(\frac{cd}{\gamma}\right) + \left(\frac{\beta+cX_*d}{\gamma}\right)\sqrt{\frac{2\ln\frac{1}{\delta}}{d_l}}.$$

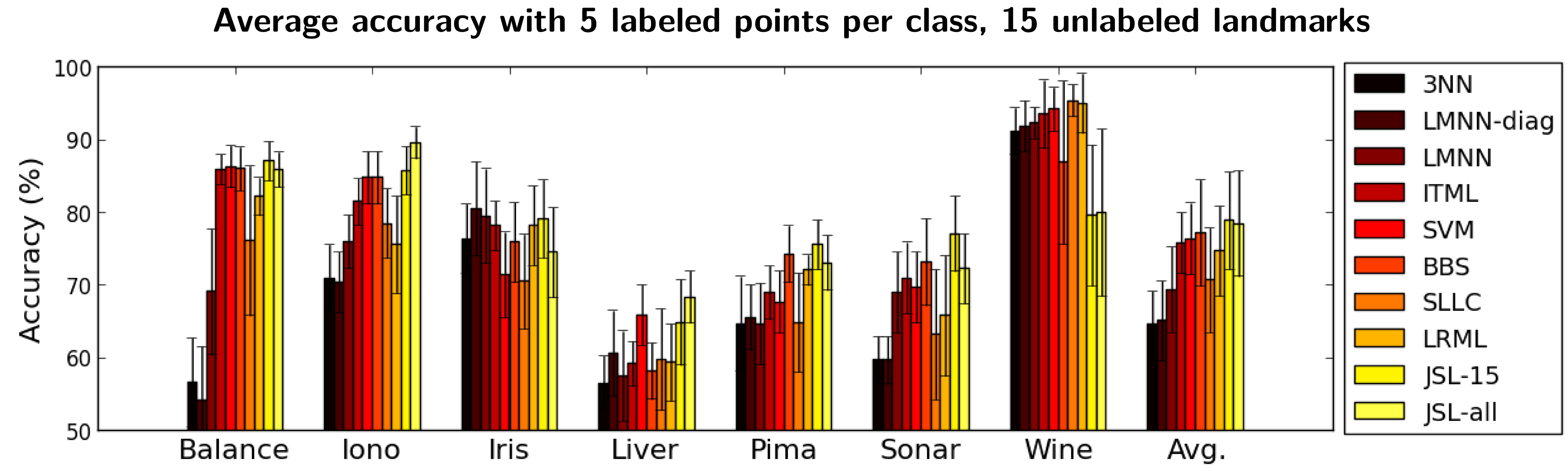
empirical risk — Rademacher complexity

Experiments

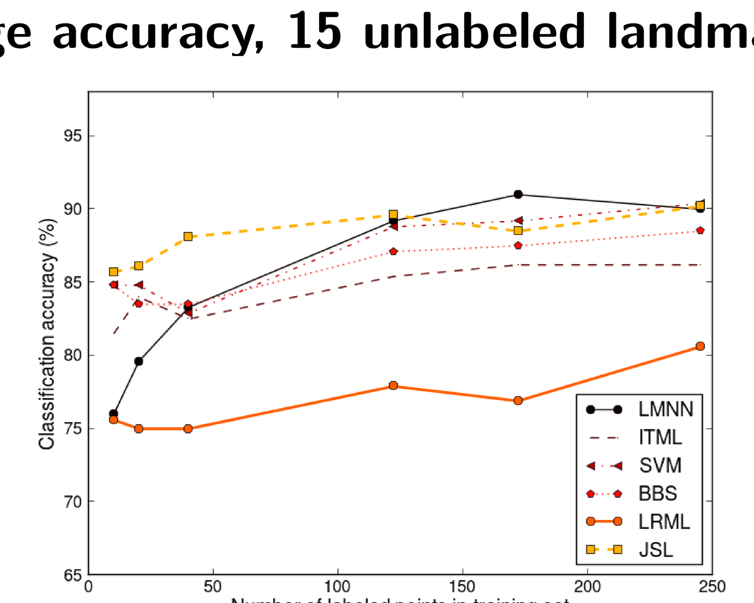
Average accuracy over all datasets

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	65.1±5.5	68.2±5.6	71.5±5.2
LMNN	69.4±5.9	70.9±5.3	73.2±5.2
ITML	75.8±4.2	76.5±4.5	76.3±4.8
LRML	74.7±6.2	75.3±5.9	75.8±5.2
SVM	76.4±4.9	76.2±7.0	77.7±6.4
BBS	77.2±7.3	77.0±6.2	77.3±6.3
SLLC	70.5±7.2	75.9±4.5	75.8±4.8
JSL-15	**78.9**±6.7	**77.6**±5.5	77.7±6.4
JSL-all	78.2±7.3	76.6±5.8	**78.4**±6.7

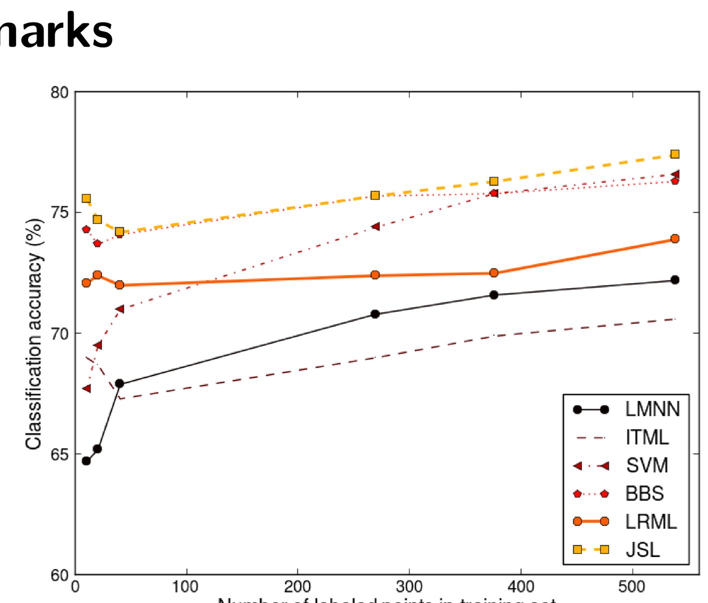
Average accuracy with 5 labeled points per class, 15 unlabeled landmarks



Average accuracy, 15 unlabeled landmarks



(a) Ionosphere



(b) Pima

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 *COLT*, pages 287–298. Omnipress, 2008.