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

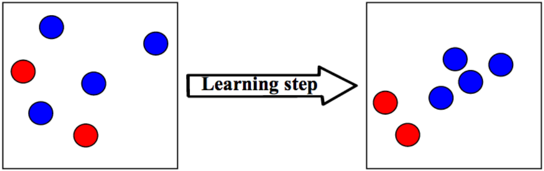
[2]Université Grenoble Alpes, CNRS-LIG/AMA, France

ECML PKDD 2015

Metric Learning

Metric Learning [Yan06, BHS13]

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



Mahalanobis Distance Learning

Find the positive semi-definite (PSD) matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ parameterizing a Mahalanobis distance

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

such that $d_{\mathbf{A}}^2$ best satisfies the constraints.

Limitations

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

Solution

- Optimize similarity function $K : \mathcal{X} \times \mathcal{X} \rightarrow [-1, 1]$ instead of distances.
- Consistency guarantees on K.
- Generalization guarantees on the classifier using K.

(ϵ, γ, τ)-Good Framework

(ϵ, γ, τ)-Good Similarity Functions

Some of the first results on how the properties of the **similarity function** influence its performance in **linear classification**.

Definition

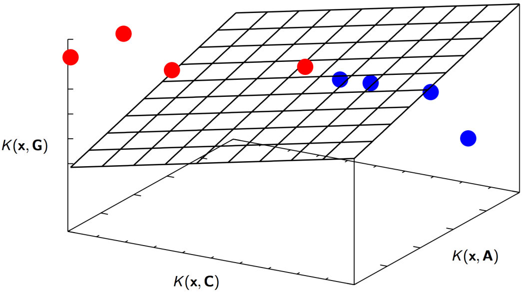
[BBS08] $K : \mathcal{X} \times \mathcal{X} \to [-1, 1]$ is a (ϵ, γ, τ)-good similarity function in hinge loss for a learning problem P if there exists a random indicator function $R(\mathbf{x})$ defining a probabilistic set of "landmarks" such that the following conditions hold:

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

Learning with (ϵ, γ, τ)-Good Similarity Functions

Theorem

[BBS08] Given K is (ϵ, γ, τ)-good, there exists a linear separator α in the projection space that has error close to ϵ at margin γ.



Learning the Classifier

Linear program

$$\min_{\boldsymbol{\alpha}} \left\{ \sum_{i=1}^{d_l} \left[1 - \sum_{j=1}^{d_u} \alpha_j y_i K(\mathbf{x}_i, \mathbf{x}_j) \right]_+ : \sum_{j=1}^{d_u} |\alpha_j| \leq 1/\gamma \right\}$$

Advantages:

- Sparsity induced by γ;
- Theoretical guarantees on $\boldsymbol{\alpha}$.

Main limitation:

- No given method to find the suited similarity function.
  - Recent work optimizing the goodness of K [BHS12].

Our contribution

- Learn both $\boldsymbol{\alpha}$ and K at the same time.
- Take advantage of unlabeled data to improve goodness of K.

Joint Similarity and Classifier Learning

Joint Similarity and Classifier Learning

Objective

We want to jointly optimize $\boldsymbol{\alpha}$ and $K_{\mathbf{A}}$ in the (ϵ, γ, τ)-good framework.

Learning Setting

- Labeled data: $\mathcal{S} = \{\mathbf{z}_i = (\mathbf{x}_i, y_i)\}^{d_l}$
- Unlabeled data: $\{\mathbf{x}_j\}^{d_u}$
- Similarity function $K_{\mathbf{A}}$, parameterized by non PSD matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$
- Instantaneous loss at point $(\mathbf{x}_i, y_i)$:
  $\ell(\mathbf{A}, \boldsymbol{\alpha}, \mathbf{z}_i = (\mathbf{x}_i, y_i))) = \left[1 - \sum_{j=1}^{d_u} \alpha_j y_i K_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j)\right]_+$

Formulation

Joint Similarity Learning (JSL)

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

$$\text{s.t.} \quad \sum_{j=1}^{d_u} |\alpha_j| \leq 1/\gamma$$

- Semi-supervised setting;
- Averaged constraints;
- Generic form of similarity and regularization;
- Solved by alternating optimization steps over $\boldsymbol{\alpha}$ and $\mathbf{A}$.

Choice of Similarity and Regularization

Similarity Functions

- $K^1_{\mathbf{A}}(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{A} \mathbf{x}'$

- $K^2_{\mathbf{A}}(\mathbf{x}, \mathbf{x}') = 1 - (\mathbf{x} - \mathbf{x}')^T \mathbf{A} (\mathbf{x} - \mathbf{x}')$

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

Theoretical Analysis

Theoretical Analysis

We want to bound the **goodness in generalization** of our learned similarity and classifier:

$$\mathcal{E}(\mathbf{A}, \boldsymbol{\alpha}) = \mathbb{E}_{\mathbf{z} \sim \mathcal{Z}} \ell(\mathbf{A}, \boldsymbol{\alpha}, \mathbf{z})$$

by the **empirical goodness**:

$$\mathcal{E}_{\mathcal{S}}(\mathbf{A}, \boldsymbol{\alpha}) = \frac{1}{d_l} \sum_{i=1}^{d_l} \ell(\mathbf{A}, \boldsymbol{\alpha}, \mathbf{z}_i).$$

Theoretical frameworks

- Uniform stability [BE02]
- Algorithmic robustness [XM12]
- VC dimension, Rademacher complexity and other similar.

Rademacher Complexity

Rademacher average over $\mathcal{F}$

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

Rademacher complexity

$$\mathcal{R}_n(\mathcal{F}) := \mathbb{E}_{\mathcal{S}} \hat{\mathcal{R}}_{\mathcal{S}}(\mathcal{F}), \forall n$$

where

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

(β, c)-Admissibility

Definition

A pairwise similarity function $K_{\mathbf{A}} : \mathcal{X} \times \mathcal{X} \to [-1, 1]$, parameterized by a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, is said to be (β, c)-admissible if, for any matrix norm $|| \cdot ||$, there exist $\beta, c \in \mathbb{R}$ such that

$$\forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}, |K_{\mathbf{A}}(\mathbf{x}, \mathbf{x}')| \leq \beta + c \cdot ||\mathbf{x}'\mathbf{x}^T|| \cdot ||\mathbf{A}||.$$

Examples

- $K^1_{\mathbf{A}}(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{A} \mathbf{x}'$ 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.

Bounding True Risk with Rademacher Complexity

Theorem (Generalization bound)

Let $(\mathbf{A}_\mathcal{S}, \boldsymbol{\alpha}_\mathcal{S})$ 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 $K_\mathbf{A}$* $X_* = \sup_{\mathbf{x},\mathbf{x}' \in \mathcal{X}} \|\mathbf{x}'\mathbf{x}^T\|_*$

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

empirical risk *Rademacher complexity*

- Convergence rate in $\mathcal{O}\left(\frac{1}{\sqrt{d_l}}\right)$.

Experiments

Experimental Setup

Methods:

1. *Linear classifiers*
   - Linear SVM with L_2 regularization;
   - BBS [BBS08];
   - SLLC [BHS12];
   - **JSL**;
2. *Nearest neighbor approaches*
   - 3NN – euclidean distance;
   - ITML [DKJ+07];
   - LMNN and LMNN-diag [WS08, WS09];
   - LRML [HLC10], semi-supervised setting.

Settings:

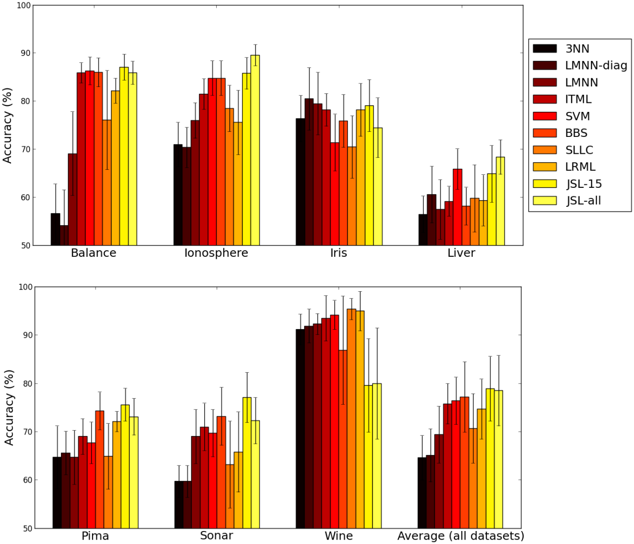
- Small quantities of labeled data: 5, 10, 20 examples per class;
- 15 unlabeled examples, or the whole training set.

Datasets:

	Balance	Ionosphere	Iris	Liver	Pima	Sonar	Wine
# Instances	625	351	150	345	768	208	178
# Dimensions	4	34	4	6	8	60	13
# Classes	3	2	3	2	2	2	3

Accuracy Comparison

5 labeled points per class

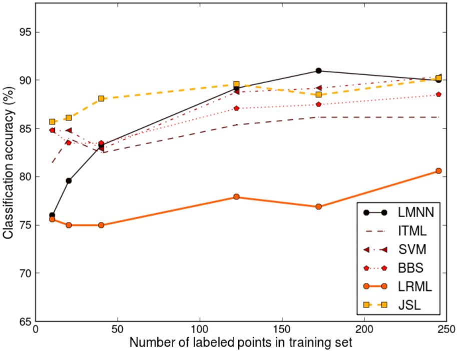


Overall Accuracy Comparison

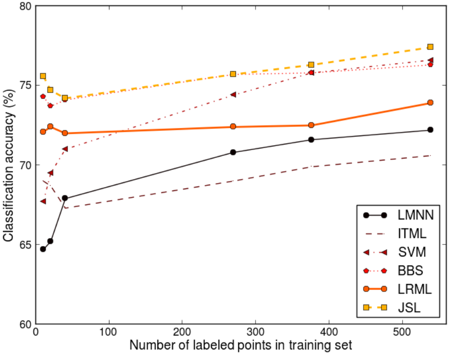
Method	5 pts./cl.	10 pts./cl.	20 pts./cl.
3NN	64.6±4.6	68.5±5.4	70.4±5.0
LMNN-diag	65.1±5.5	68.2±5.6	71.5±5.2
LMNN	69.4±5.9	70.9±5.3	73.2±5.2
ITML	75.8±4.2	76.5±4.5	76.3±4.8
SVM	76.4±4.9	76.2±7.0	77.7±6.4
BBS	77.2±7.3	77.0±6.2	77.3±6.3
SLLC	70.5±7.2	75.9±4.5	75.8±4.8
LRML	74.7±6.2	75.3±5.9	75.8±5.2
JSL-15	**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

Impact of the amout of labeled data

15 unlabeled landmarks



(a) Ionosphere



(b) Pima

Conclusion

Conclusion

- New **semi-supervised** metric learning framework;
- **Joint learning** of a metric and a global separator;
- General similarity function and regularizer;
- **Theoretical guarantees** using Rademacher complexity.

Future work

- Bigger datasets $\rightarrow$ online algorithm;
- Landmarks selection heuristiques.

Thank you!

Come see the poster!

References I

Acknowledgements

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

Maria-Florina Balcan, Avrim Blum, and Nathan Srebro.
Improved guarantees for learning via similarity functions.
In *COLT*, pages 287–298. Omnipress, 2008.

Olivier Bousquet and André Elisseeff.
Stability and generalization.
JMLR, 2:499–526, March 2002.

Aurélien Bellet, Amaury Habrard, and Marc Sebban.
Similarity learning for provably accurate sparse linear classification.
In *ICML*, pages 1871–1878, 2012.

Aurélien Bellet, Amaury Habrard, and Marc Sebban.
A survey on metric learning for feature vectors and structured data.
Technical report, 2013.

Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon.
Information-theoretic metric learning.
In *ICML*, pages 209–216, New York, NY, USA, 2007. ACM.

Steven C. H. Hoi, Wei Liu, and Shih-Fu Chang.
Semi-supervised distance metric learning for collaborative image retrieval and clustering.
TOMCCAP, 6(3), 2010.

References II

K.Q. Weinberger and L.K. Saul.
Fast solvers and efficient implementations for distance metric learning.
In *ICML*, pages 1160–1167. ACM, 2008.

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

Huan Xu and Shie Mannor.
Robustness and generalization.
Machine Learning, 86(3):391–423, 2012.

Liu Yang.
Distance metric learning: A comprehensive survey, 2006.