Learning Similarities for Linear Classification: Theoretical Foundations and Algorithms

Maria-Irina Nicolae

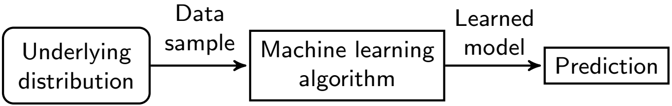
Laboratoire Hubert Curien, Université de Saint-Etienne, Université de Lyon
Laboratoire d'Informatique de Grenoble, Université Grenoble Alpes

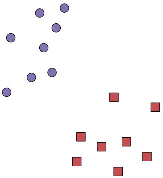
Reviewers: Ludovic Denoyer (UPMC-LIP6), Antoine Cornuéjols (AgroParisTech)
Examiners: Maria-Florina Balcan (CMU), Amaury Habrard (U. St-Etienne), Liva Ralaivola (U. Aix-Marseille)
Supervisors: Éric Gaussier (U. Grenoble-Alpes) and Marc Sebban (U. St-Etienne)

PhD Defense, December 2, 2016

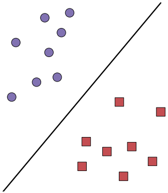
Scientific Context

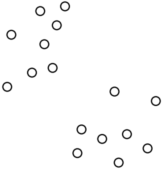
Machine Learning



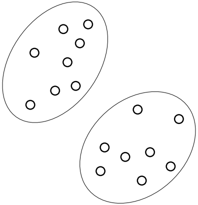


Supervised learning

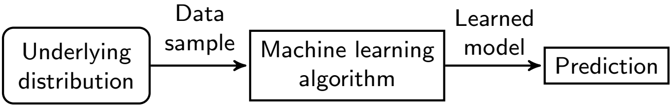


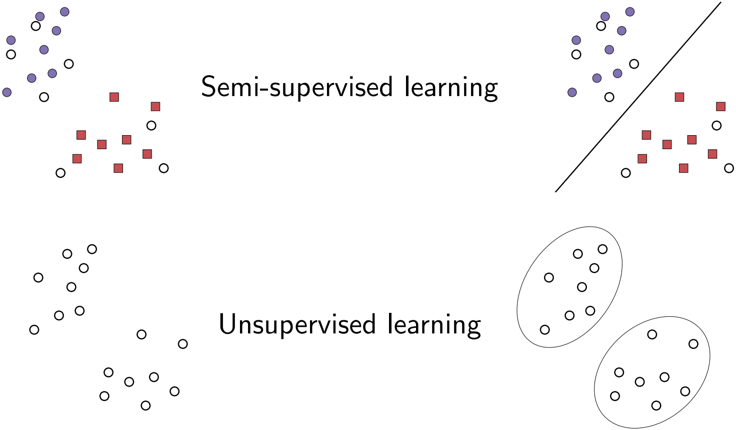


Unsupervised learning



Machine Learning

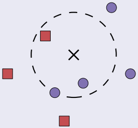




Metric Learning

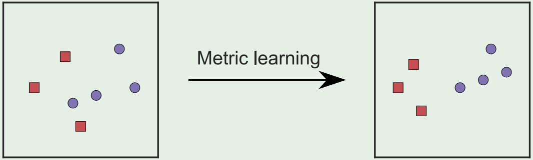
Performance depends on metrics

Most machine learning algorithms compare examples using a metric.



Adapting the metric to the problem

Learn a custom metric which better discriminates the examples.



Typically, it boils down to learning a new representation space.

Representation Learning

Importance

Representation learning is often key in machine learning algorithms.

Subfields of representation learning

- Kernel learning;
- Multiple kernel learning;
- Dictionary learning;
- Deep learning;
- Metric learning.

Differences come from:

- Orthogonality of the features;
- Implicit vs. explicit representation;
- Sparse vs. infinite representation space;
- Parametric vs. non parametric.

Background and Preliminaries

Input

A **sample** of n labeled examples $\mathcal{S} = \{z_i = (\mathbf{x}_i, y_i)\}_{i=1}^n$ independent and identically distributed (i.i.d.) according to an unknown distribution P.

Output

A model $h \in H$ that **best predicts** the labels of unseen examples.

Definition (True risk)

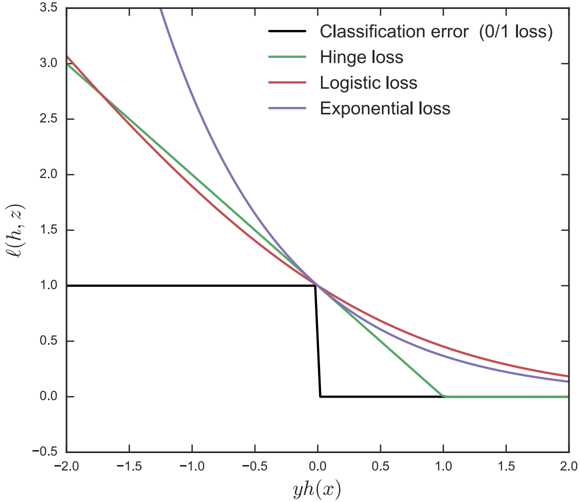
The expected loss suffered by h on the distribution P:
$R_P^\ell(h) = \mathbb{E}_{z \sim P}[\ell(h, z)]$.

Definition (Empirical risk)

The average loss incurred by h on the sample $\mathcal{S}$: $R_\mathcal{S}^\ell(h) = \frac{1}{n}\sum_{i=1}^n \ell(h, z_i)$.

Supervised Learning

Loss Functions for Binary Classification



Minimize an objective function of the form:

$$h_{\mathcal{S}} = \arg\min_{h \in H} R^{\ell}_{\mathcal{S}}(h) + \lambda ||h||.$$

PAC bounds

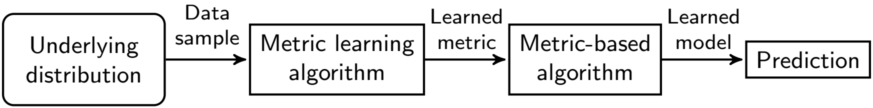
Bound the deviation of the empirical risk from the true risk of a hypothesis h, i.e. its **capacity to generalize** to an unseen sample:

$$|R^{\ell}_{P}(h) - R^{\ell}_{\mathcal{S}}(h)| \leq \mathcal{O}(complexity(h)/\sqrt{n}).$$

Theoretical frameworks

- Uniform convergence using VC dimension [VC71], Rademacher complexity [BM03] and other similar,
- Uniform stability [BE02],
- Algorithmic robustness [XM12].

Metric Learning



Mahalanobis distance learning

Find the positive semi-definite (PSD) matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$ parameterizing a Mahalanobis distance $d^2_{\mathbf{M}}$ to best satisfy some constraints:

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

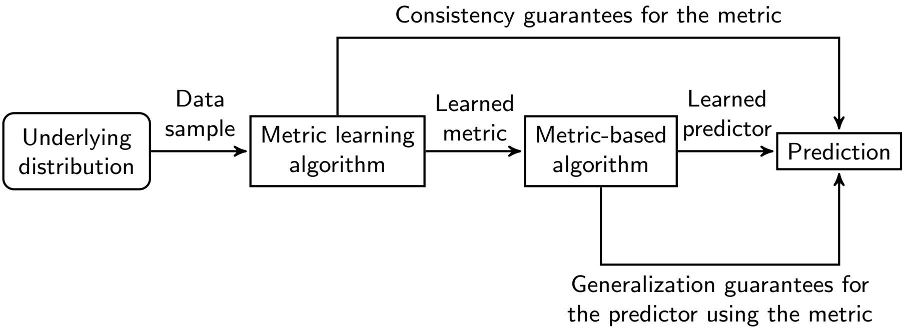
Similarity judgments (constraints)

- Pair-based: $\mathbf{x}$ and $\mathbf{x}'$ are (dis)similar;
- Triplet-based: $\mathbf{x}$ is more similar to $\mathbf{x}'$ than to $\mathbf{x}''$.

Existing methods differ in the choice of the **metric**, the **constraints**, the **loss function** and the **regularizer**.

Generalization Guarantees in Metric Learning

The question of the generalization capacity can be asked at two levels: the **metric** and the **predictor** using it.



Only a few methods provide generalization guarantees for:

- the **learned metric** $d_{\mathbf{M}}$ itself [JWZ09, BT11, CGY12];
- the **performance** of the algorithm using it [BBS08].

(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**.

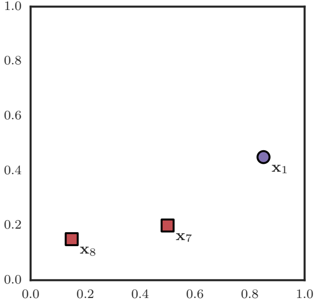
Definition

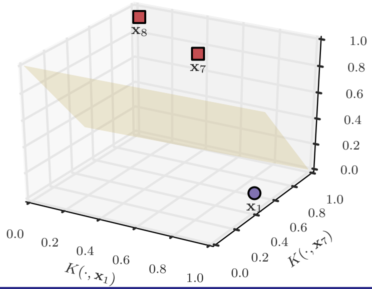
[BBS08] $K \in [-1, 1]$ is an (ϵ, γ, τ)-good similarity function in hinge loss for a learning problem P if there exists a random indicator function $R(\mathbf{x})$ defining a probabilistic set of landmarks such that the following conditions hold:

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. $\Pr_{\mathbf{x}'}(R(\mathbf{x}')) \geq \tau$. $\qquad \epsilon, \gamma, \tau \in [0, 1]$

[BBS08] Maria-Florina Balcan, Avrim Blum, and Nathan Srebro. Improved Guarantees for Learning via Similarity Functions. In *COLT*, 2008.

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





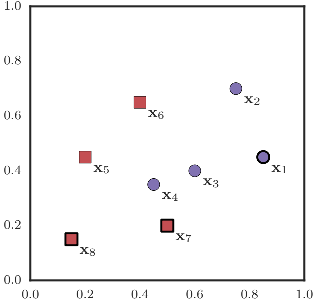
Theorem

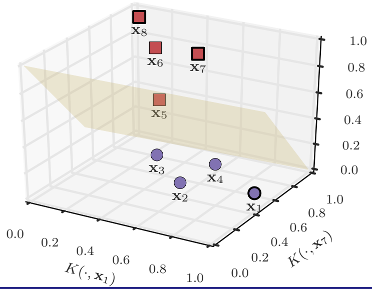
[BBS08] Given K is (ϵ, γ, τ)-good, there exists a linear separator α in the projection space that has error close to ϵ at margin γ.

Linear program (BBS)

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

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.

Linear program (BBS)

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

Summary

Advantages

- Theoretical guarantees on α;
- Semi-supervised framework.

Limitations

- Standard similarity functions might poorly satisfy the definition;
- No given method to find the suited similarity function.

Solution

- Learn the similarity function from a data sample;
- Directly optimize its **empirical goodness**;
- This implies **guarantees** for the linear classifier.

Contributions of the Thesis

- Optimize **similarity functions** instead of distances:
  - Less costly;
  - Gives access to a larger class of functions.
- Optimize the **goodness** of the similarities $\rightarrow$ generalization guarantees for the algorithm using the similarity.
- Derive **consistency guarantees** for the learned similarity.

More precisely

- Joint similarity and classifier learning for feature vectors.
- Similarity learning for multivariate time series classification.

Previous results using the (ϵ, γ, τ)-good framework

- String edit distance learning [BHS11];
- Learning a bilinear similarity for feature vectors [BHS12].

Joint Similarity and Classifier Learning for Feature Vectors

Optimizing the (ϵ, γ, τ)-Goodness in Metric Learning

(ϵ, γ, τ)-Goodness Criterion

$$\mathbb{E}_{(\mathbf{x},y)\sim P}\left[\left[1 - y\mathbb{E}_{(\mathbf{x}',y'),R(\mathbf{x}')}\left[yy'K_{\mathbf{M}}(\mathbf{x},\mathbf{x}'))|R(\mathbf{x}')\right]/\gamma\right]_+\right] \leq \epsilon.$$

To be satisfied on average.

Linear program (BBS)

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

Landmarks labels are not used.

Problem Formulation

Joint Similarity Learning (JSL)

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

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

Properties

- Semi-supervised setting $\rightarrow$ can use a small quantity of labeled data;
- Averaged constraints;
- Generic form of similarity and regularization;
- Convex for a large range of similarities and regularizers;
- Solved by alternating optimization steps over $\boldsymbol{\alpha}$ and $\mathbf{M}$.

Choice of Similarity and Regularization

Similarity functions

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

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

- 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 using Rademacher Complexity

Rademacher average over $\mathcal{F}$

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

Rademacher complexity

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

where

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

Bounding True Risk with Rademacher Complexity

Definition ((β, c)-admissibility)

A similarity function $K_{\mathbf{M}} : \mathcal{X} \times \mathcal{X} \to [-1, 1]$ parameterized by $\mathbf{M} \in \mathbb{R}^{d \times d}$ is (β, c)-admissible if, for any matrix norm $|| \cdot ||$, there exist $\beta, c \in \mathbb{R}$ s.t. $\forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}, |K_{\mathbf{M}}(\mathbf{x}, \mathbf{x}')| \leq \beta + c \cdot ||\mathbf{x}'\mathbf{x}^T|| \cdot ||\mathbf{M}||$.

Theorem (Generalization bound)

Let $(\mathbf{M}_{\mathcal{S}}, \boldsymbol{\alpha}_{\mathcal{S}})$ be the solution to JSL and $K_{\mathbf{M}}$ 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 $K_{\mathbf{M}}$* $X_* = \sup_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}} ||\mathbf{x}'\mathbf{x}^T||_*$

$$|R_P^{\ell}(\mathbf{M}_{\mathcal{S}}, \boldsymbol{\alpha}_{\mathcal{S}}) - R_{\mathcal{S}}^{\ell}(\mathbf{M}_{\mathcal{S}}, \boldsymbol{\alpha}_{\mathcal{S}})| \leq 4\mathfrak{R}_m \left(\frac{cd}{\gamma}\right) + \left(\frac{\beta + cX_* d}{\gamma}\right) \sqrt{\frac{2 \ln \frac{1}{\delta}}{m}}.$$

empirical risk *Rademacher complexity* *convergence in $\mathcal{O}\left(\frac{1}{\sqrt{m}}\right)$*

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.

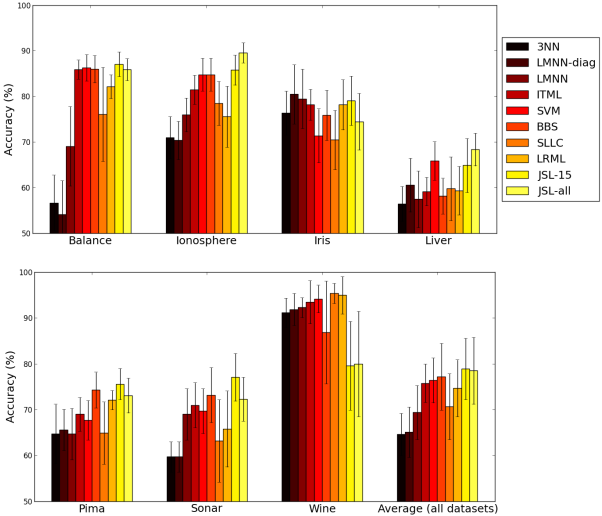
Settings:

- 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

5 labeled points per class



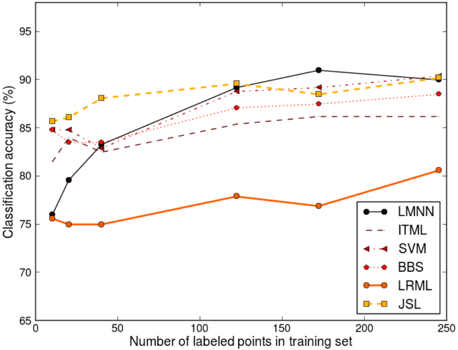
Experimental Results

Overall Accuracy Comparison

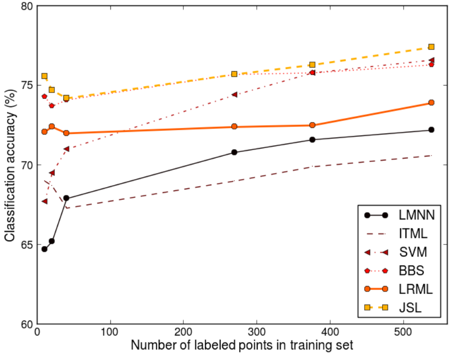
Method	5 pts./cl.	10 pts./cl.	20 pts./cl.
3NN	64.6±4.6	68.5±5.4	70.4±5.0
LMNN-diag	65.1±5.5	68.2±5.6	71.5±5.2
LMNN	69.4±5.9	70.9±5.3	73.2±5.2
ITML	75.8±4.2	76.5±4.5	76.3±4.8
SVM	76.4±4.9	76.2±7.0	77.7±6.4
BBS	77.2±7.3	77.0±6.2	77.3±6.3
SLLC	70.5±7.2	75.9±4.5	75.8±4.8
LRML	74.7±6.2	75.3±5.9	75.8±5.2
JSL-15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

Experimental Results
Impact of the Amount of Labeled Data

15 unlabeled landmarks



(a) Ionosphere



(b) Pima

Summary of JSL

- 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 and algorithmic robustness.

Publications

- M.-I. Nicolae, É. Gaussier, A. Habrard, and M. Sebban. Joint semi-supervised similarity learning for linear classification. In *ECML/PKDD*, 2015a.
- M.-I. Nicolae, M. Sebban, A. Habrard, É. Gaussier, and M.-R. Amini. Algorithmic Robustness for Semi-Supervised (ϵ, γ, τ)-Good Metric Learning. In *ICONIP*, pages 253–263, 2015b.
- M.-I. Nicolae, M. Sebban, A. Habrard, É. Gaussier, and M.-R. Amini. Algorithmic Robustness for Learning via (ϵ, γ, τ)-Good Similarity Functions. In *ICLR Workshop*, 2015.

Learning Similarities for Time Series Classification

Motivation

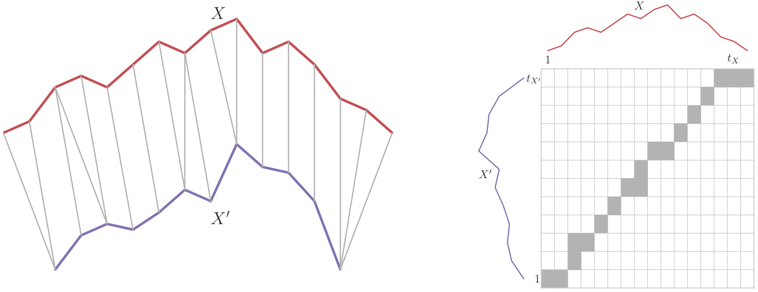
Time series

Vast presence of time series in real-world applications.

Metric learning for time series

- Little work in this field.
- Most of it focused on adapting known methods to the univariate case.

Dynamic Time Warping



Find the optimal alignment between two time series based on a cost matrix:

- **Quadratic complexity** in the length of the time series;
- Univariate case: often Euclidean distance;
- Multivariate case: need a measure for comparing time moments with multiple features.

Bilinear Similarity for Time Series

Time series **alignment** of length $t_{\mathbf{AB}}$ using DTW:

$$\mathbf{Y}_{\mathbf{AB}} = \mathrm{DTW}(\mathbf{A}, \mathbf{B}).$$

Affinity for aligning time moments $0 < i \leq t_{\mathbf{A}}$ and $0 < j \leq t_{\mathbf{B}}$ between series **A** and **B**:

$$\mathbf{C}_{\mathbf{M}}(\mathbf{A}, \mathbf{B})_{i,j} = \mathbf{a}_i^T \cdot \mathbf{M} \cdot \mathbf{b}_j.$$

Affinity matrix under metric $\mathbf{M} \in \mathbb{R}^{d \times d}$ for computing the cost of alignment:

$$\mathbf{C}_{\mathbf{M}}(\mathbf{A}, \mathbf{B}) = \mathbf{A} \cdot \mathbf{M} \cdot \mathbf{B}^T.$$

Bilinear similarity

Let $K_{\mathbf{M}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ of form:

$$K_{\mathbf{M}}(\mathbf{A}, \mathbf{B}) = \mathrm{Tr}(\mathbf{C}_{\mathbf{M}}(\mathbf{A}, \mathbf{B})^T \cdot \mathbf{Y}_{\mathbf{AB}})/t_{\mathbf{AB}}.$$

Computes the score for aligning **A** and **B** under metric **M**.

Learning the Similarity

Improve the (ϵ, γ, τ)-goodness of $K_{\mathbf{M}}$:

$$\mathbb{E}_{(\mathbf{A},y)}\left[\left[1 - \mathbb{E}_{(\mathbf{B},y'),R(\mathbf{B})}\left[yy'K_{\mathbf{M}}(\mathbf{A},\mathbf{B}))|R(\mathbf{B})\right]/\gamma\right]_{+}\right] \leq \epsilon.$$

But we do not have access to expected values.

Similarity Learning for Time Series (SLTS)

Optimize the empirical value of the goodness criterion over sample $\mathcal{S}$ w.r.t. the set of landmarks $\mathcal{L}$:

$$\min_{\mathbf{M}} \frac{1}{m} \sum_{(\mathbf{A},y)\in\mathcal{S}} \left[1 - \frac{1}{n\gamma} \sum_{(\mathbf{B},y')\in\mathcal{L}} yy'K_{\mathbf{M}}(\mathbf{A},\mathbf{B})\right]_{+} + \lambda||\mathbf{M}||_{\mathcal{F}}^2.$$

Properties

- Convex formulation;
- Based on landmarks $\rightarrow$ does not need to compute DTW and the similarity for all pairs.

An algorithm is **stable** if its output is robust to small changes in its input. Uniform stability allows the derivation of generalization bounds.

Lemma

Given a training sample $\mathcal{S}$ of m examples drawn i.i.d. from P, our algorithm SLTS has uniform stability in κ/m with $\kappa = \frac{4d}{\gamma^2\lambda}$, that is:

$$\sup_{(\mathbf{A},l)\sim P} |\ell(\mathbf{M},(\mathbf{A},l)) - \ell(\mathbf{M}^i,(\mathbf{A},l))| \leq \frac{\kappa}{m},$$

where $\mathbf{M}^i$ is obtained by learning on $\mathcal{S}$ after replacing the ith example with a new one.

Bounding True Risk with Uniform Stability

Theorem (Generalization bound)

For any $0 < \delta < 1$, with probability $1 - \delta$, for any matrix $\mathbf{M}$ learned with SLTS, we have:

true risk — *# features*

$$|R_P^\ell(\mathbf{M}) - R_S^\ell(\mathbf{M})| \leq \frac{4d}{\gamma^2 \lambda m} + \left(\frac{4d}{\gamma^2 \lambda} + \frac{1}{\gamma}\sqrt{\frac{2d}{\lambda}} \right) \sqrt{\frac{2 \log \frac{2}{\delta}}{m}}.$$

empirical risk — *margin* — *convergence in $\mathcal{O}\left(\frac{1}{\sqrt{m}}\right)$*

- Independence from the length of the time series and the alignments.

Experimental Setup

Methods:

Nearest neighbor approaches:

- 1NN
- LDMLT

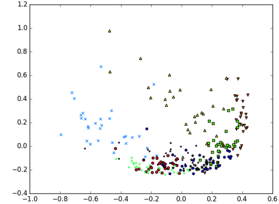
Linear classifiers:

- L_2 regularized SVM
- BBS
- **SLTS**.

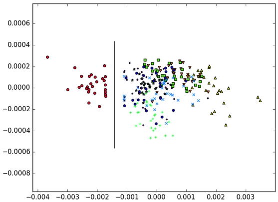
UCI Datasets [Lic13]:

Dataset	#Instances	Length	#Feat.	#Classes
Japanese vowels	640	7-29	12	9
Auslan	675	47-95	22	25
Arabic digits	8800	4-93	13	10
Robot execution failure				
LP1	88	15	6	4
LP2	47	15	6	5
LP3	47	15	6	4
LP4	117	15	6	3
LP5	164	15	6	5

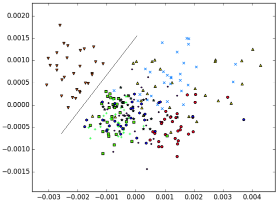
2D PCA on Japanese Vowels



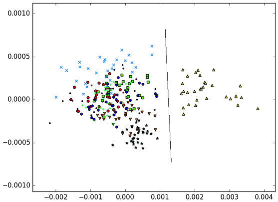
(a) No metric learning



(b) Metric for class 1



(c) Metric for class 2



(d) Metric for class 3

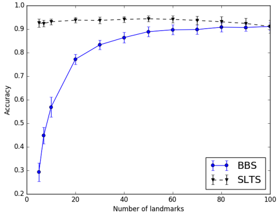
Experimental Results

Classification Accuracy (%)

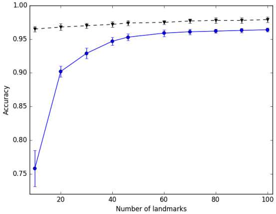
Method	Japanese vowels	Auslan	Arabic digits	Robot failure	Avg.
1NN	93.8	77.8±2.1	94.7	68.8±7.5	92.1
LDMLT	97.3	95.0±1.3	96.9	71.9±7.0	95.6
L_2 SVM	97.8±0.1	92.6±0.1	93.3±0.0	60.6±6.5	92.2
BBS	97.1±0.5	91.1±1.6	96.4±0.3	66.9±10.6	94.7
SLTS	97.1±0.4	91.1±2.7	97.9±0.4	67.0±7.8	**95.8**

- SLTS has comparable performance to the other methods;
- It has **theoretical guarantees** and does not need to compute all the alignments.

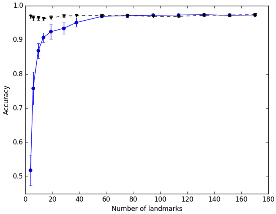
Classification accuracy for SLTS and BBS



(a) Auslan



(b) Arabic digits



(c) Japanese vowels

Summary of SLTS

- Novel method for **learning similarities** for multivariate time series classification.
- Metric consistency based on **uniform stability**.
- First method with **theoretical guarantees** for time series.

Publications

- M.-I. Nicolae, É. Gaussier, A. Habrard, and M. Sebban. Similarity Learning for Time Series Classification. Technical report, University of Saint-Etienne, 2016. arXiv:1610.04783. To be submitted to the journal track of *ECML/PKDD* 2017 and *MLJ*.

General Perspectives

General Perspectives

- Introduce nonlinearity by learning multiple local metrics.

General Perspectives

- Introduce nonlinearity by learning multiple local metrics.
- Challenge the learning pairs.

General Perspectives

- Introduce nonlinearity by learning multiple local metrics.
- Challenge the learning pairs.
- Goodness in similarity learning for local classification.

General Perspectives

- Introduce nonlinearity by learning multiple local metrics.
- Challenge the learning pairs.
- Goodness in similarity learning for local classification.
- Metric learning for an unsupervised setting with generalization guarantees.

References I

Maria-Florina Balcan, Avrim Blum, and Nathan Srebro.
Improved Guarantees for Learning via Similarity Functions.
In Rocco A. Servedio and Tong Zhang, editors, *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, pages 287–298. Omnipress, 2008.

Olivier Bousquet and André Elisseeff.
Stability and Generalization.
Journal of Machine Learning Research (JMLR), 2:499–526, March 2002.

Aurélien Bellet, Amaury Habrard, and Marc Sebban.
Learning Good Edit Similarities with Generalization Guarantees.
In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, pages 188–203, 2011.

Aurélien Bellet, Amaury Habrard, and Marc Sebban.
Similarity Learning for Provably Accurate Sparse Linear Classification.
In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1871–1878, 2012.

Peter L. Bartlett and Shahar Mendelson.
Rademacher and Gaussian Complexities: Risk Bounds and Structural Results.
Journal of Machine Learning Research (JMLR), 3:463–482, March 2003.

Wei Bian and Dacheng Tao.
Learning a Distance Metric by Empirical Loss Minimization.
In Toby Walsh, editor, *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1186–1191. IJCAI/AAAI, 2011.

References II

Qiong Cao, Zheng-Chu Guo, and Yiming Ying.
Generalization Bounds for Metric and Similarity Learning.
ArXiv e-prints, arXiv:1207.5437, 2012.

Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon.
Information-Theoretic Metric Learning.
In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, ICML '07, 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.
ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 6(3):1–26, aug 2010.

Rong Jin, Shijun Wang, and Yang Zhou.
Regularized Distance Metric Learning: Theory and Algorithm.
In Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 862–870. Curran Associates, Inc., 2009.

Moshe Lichman.
UCI Machine Learning Repository, 2013.

Vladimir N. Vapnik and Alexey Y. Chervonenkis.
On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities.
Theory of Probability and Its Applications (TPA), 16(2):264–280, 1971.

Kilian Q. Weinberger and Lawrence K. Saul.
Fast Solvers and Efficient Implementations for Distance Metric Learning.
In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 1160–1167. ACM, 2008.

References III

Kilian Q. Weinberger and Lawrence K. Saul.
Distance Metric Learning for Large Margin Nearest Neighbor Classification.
The Journal of Machine Learning Research (JMLR), 10:207–244, 2009.

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