Discriminative Robust Transformation Learning

Jiaji Huang, Qiang Qiu, Guillermo Sapiro and Robert Calderbank

Department of Electrical Engineering, Duke University, Durham, NC 27708

Abstract

- Learning Discriminative feature is a classic problem
- Yet the robustness to small training size is less studied
- Proposed: use local isometry to improve robustness and reduce over-fitting

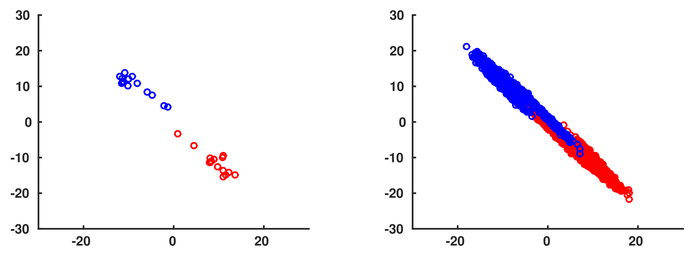
Motivating Example

A discriminative linear transform:

Let $\mathcal{P}$ be the set of all the sample pairs

$\ell_{i,j} = 1/-1$ if $\mathbf{x}_i, \mathbf{x}_j \in$ same/distinct class/classes. $t(1) < t(-1)$

$$\min_{\mathbf{A}} \sum_{(i,j)\in\mathcal{P}} \frac{1}{|\mathcal{P}|} \max\left\{0, \ell_{i,j}\left[\|\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j\| - t(\ell_{i,j})\right]\right\}$$



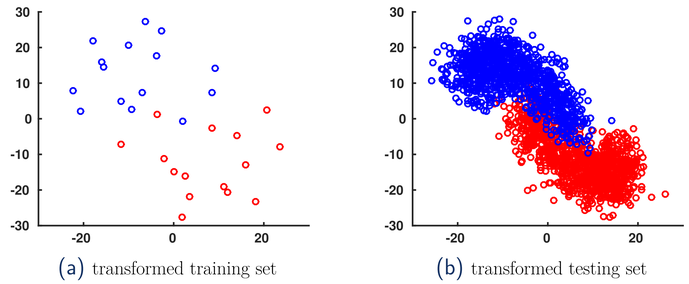
(a) transformed training set (b) transformed testing set

The discriminative linear transform applied to two noisy half-moons: (a) very discriminative on training data; (b) but mixes the classes on test data

A local structure-preserving linear transform:

Let $\mathcal{NB}$ be the set of all the local neighbors

$$\min_{\mathbf{A}} \sum_{(i,j)\in\mathcal{NB}} \frac{1}{|\mathcal{NB}|} \Big| \|\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j\| - \|\mathbf{x}_i - \mathbf{x}_j\| \Big|.$$



(a) transformed training set (b) transformed testing set

A linear transform that preserves local distance: (a) less discriminative on training data; (b) but still able to distinguish the classes on test set

Guess: Preserving local structures may reduce the training-to-testing degradation.

A General framework

- Data space :$\mathcal{X}$
- label space $\mathcal{Y} = \{1, \ldots, L\}$; $\mathcal{Z} \stackrel{\text{def}}{=} \mathcal{X} \times \mathcal{Y}$; i-th sample $\mathbf{z}_i = (\mathbf{x}_i, y_i)$.
- Training set: $\mathcal{T} \in \mathcal{X}$, $|\mathcal{T}| = N$.
- A general feature transform $f_{\boldsymbol{\alpha}}(\mathbf{x}) : \mathcal{X} \mapsto \mathcal{F}$
- Learning **D**iscriminative **R**obust **T**ransform (DRT):

$$\boldsymbol{\alpha}_{\mathcal{T}} = \arg\min_{\boldsymbol{\alpha}} \; \lambda \underbrace{\sum_{i,j\in\mathcal{P}} \frac{1}{|\mathcal{P}|} \max\left\{0, \ell_{i,j}\left[\rho\left(f_{\boldsymbol{\alpha}}(\mathbf{x}_i), f_{\boldsymbol{\alpha}}(\mathbf{x}_j)\right) - t(\ell_{i,j})\right]\right\}}_{\text{Empirical loss: } R_{emp}(\boldsymbol{\alpha})}$$
$$+ (1-\lambda) \sum_{(i,j)\in\mathcal{NB}} \frac{1}{|\mathcal{NB}|} \Big| \rho(f_{\boldsymbol{\alpha}}(\mathbf{x}_i), f_{\boldsymbol{\alpha}}(\mathbf{x}_j)) - \rho(\mathbf{x}_i, \mathbf{x}_j) \Big|$$

- Loss incurred by each pair of samples: $r_{\boldsymbol{\alpha}}(\mathbf{z}_i, \mathbf{z}_j)$.
- Expected loss: $R = \mathbb{E}_{\mathbf{z}_i,\mathbf{z}_j}[r_{\boldsymbol{\alpha}}(\mathbf{z}_i, \mathbf{z}_j)]$
- Generalization error: $R - R_{emp}$

Theoretical Analysis

Definition of (K, ϵ)-robustness [1]

A learning algorithm is (K, ϵ)-robust if $\mathcal{Z} \stackrel{\text{def}}{=} \mathcal{X} \times \mathcal{Y}$ can be partitioned into K disjoint sets $\mathcal{Z}_k, k = 1, \ldots, K$ such that for all training sets $\mathcal{T} \in \mathcal{Z}^n$, assume $\mathbf{z}_i, \mathbf{z}_j \in \mathcal{T}$, with $\mathbf{z}_i \in \mathcal{Z}_p$ and $\mathbf{z}_j \in \mathcal{Z}_q$, if $\mathbf{z}'_i \in \mathcal{Z}_p$ and $\mathbf{z}'_j \in \mathcal{Z}_q$, then

$$\left| r_{\boldsymbol{\alpha}_{\mathcal{T}}}(\mathbf{z}_i, \mathbf{z}_j) - r_{\boldsymbol{\alpha}_{\mathcal{T}}}(\mathbf{z}'_i, \mathbf{z}'_j) \right| \le \epsilon.$$

Remark: (K, ϵ) algorithms has $R - R_{emp} \le \epsilon + O(\sqrt{\frac{K}{N}})$

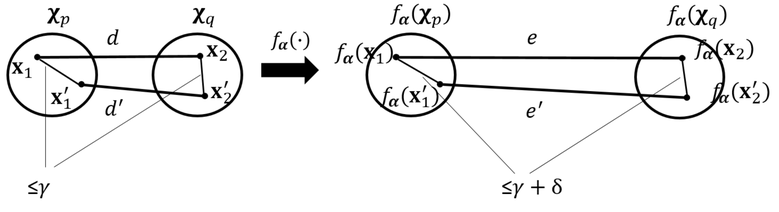
Theorem (local isometry induced robustness)

Let $f_{\boldsymbol{\alpha}}$ be a transform derived via minimizing $R_{emp}(\boldsymbol{\alpha})$ and let $\mathcal{X}_1, \ldots, \mathcal{X}_{L\mathcal{N}_{\gamma/2}(\mathcal{X},\rho)}$ be a cover of $\mathcal{X}$, where for any $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}_p$,

$$\rho(\mathbf{x}_i, \mathbf{x}_j) \le \gamma \text{ and } \mathbf{y}_i = \mathbf{y}_j$$

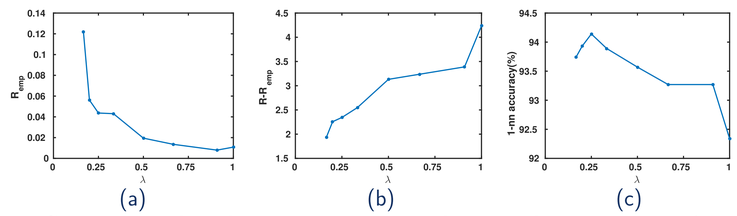
If $f_{\boldsymbol{\alpha}}$ is a δ-isometry, then it is $(L\mathcal{N}_{\gamma/2}(\mathcal{X}, \rho), 2A(\gamma + \delta))$-robust.

Proof without words



Experiments

MNIST with very small training set



(a) (b) (c)

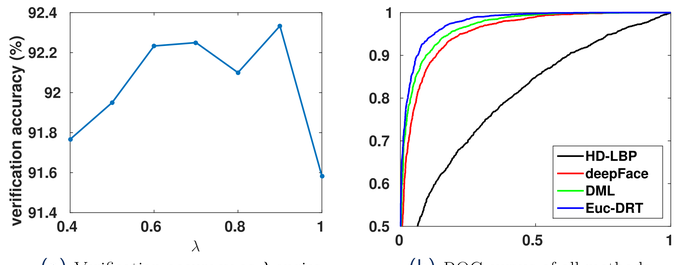
MNIST test: with only 30 training samples per class. We vary λ and assess (a) R_{emp}; (b) generalization error; and (c) 1-nn classification accuracy. Peak accuracy is achieved at $\lambda = 0.25$.

Tab1: Classification Accuracy on MNIST (Varying Training Size)

Training/class	30	50	70	100
original pixels	81.91%	86.18%	86.86%	88.49%
LeNet	87.51%	89.89%	91.24%	92.75%
DML	92.32%	94.45%	95.67%	96.19%
Euc-DRT	**94.14%**	**95.20%**	**96.05%**	**96.21%**

Face verification

- Trained $f_{\boldsymbol{\alpha}}$, a two-layer fully connected network, on WDRef
- WDRef is only with 20 subjects per class $\ll$ deepFace



(a) Verification accuracy as λ varies (b) ROC curves of all methods

Tab2: Verification accuracy and AUCs on LFW

Method	Accuracy (%)	AUC ($\times 10^{-2}$)
HD-LBP	74.73	82.22±1.00
deepFace	88.72	95.50± 0.29
DML	90.28	96.74±0.33
Euc-DRT	**92.33**	**97.77± 0.25**

Reference

[1] A. Bellet and A. Habrard. Robustness and generalization for metric learning. Neurocomputing, 151 (2015): 259-267