On the Usefulness of Similarity based Projection Spaces for Transfer Learning

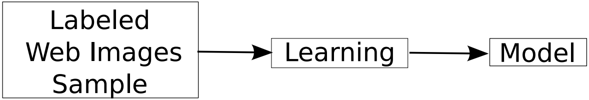
Emilie Morvant[1] Amaury Habrard[2] Stéphane Ayache[1]

[1] **Laboratoire d'Informatique Fondamentale**
QARMA Team
Aix-Marseille University, France
{emilie.morvant,stephane.ayache}@lif.univ-mrs.fr

[2] **Laboratoire Hubert Curien**
Saint-Étienne University, France
amaury.habrard@univ-st-etienne.fr

SIMBAD Workshop 2011, Venice, Italy
28th September 2011

Introduction and Motivation



Images classification : Is there a “Person” ?

Available labeled data : Images from a **Web** corpus

- Supervised Classification task
  - Test data : Images from the same **Web** corpus

  $\Rightarrow$ Low-error classifier on test data coming from the same corpus

- Domain Adaptation task
  - New test data : Images from a different **Video** corpus

  $\Rightarrow$ The classifier quality is no more guaranteed

Introduction and Motivation

A Transfer Learning Task: Domain Adaptation (DA)

Learning with Good Similarity Functions (SF)

Modifying the Projection Space for Domain Adaptation

Experimentations

Extended Work: A Little Bit of Theory

Conclusion and Perspectives

Outline

A Transfer Learning Task: Domain Adaptation (DA)

Learning with Good Similarity Functions (SF)

Modifying the Projection Space for Domain Adaptation
- An Heuristic Normalization
- An Additional Regularization Term

Experimentations

Extended Work: A Little Bit of Theory

Conclusion and Perspectives

Outline

A Transfer Learning Task: Domain Adaptation (DA)

Learning with Good Similarity Functions (SF)

Modifying the Projection Space for Domain Adaptation
- An Heuristic Normalization
- An Additional Regularization Term

Experimentations

Extended Work: A Little Bit of Theory

Conclusion and Perspectives

A Transfer Learning Task: Domain Adaptation
Formalisation

Notations

- X input space, $Y = \{-1, 1\}$ label set
- P_S **source** domain : distribution over $X \times Y$
  D_S marginal distribution over X
- P_T **target** domain : different distribution over $X \times Y$
  D_T marginal distribution over X

Expected error of an hypothesis $h : X \to Y$

- $\mathrm{err}_S(h) = \mathbb{E}_{(\mathbf{x},y)\sim P_S} \left[h(\mathbf{x}) \neq y\right]$ **source** domain error
- $\mathrm{err}_T(h) = \mathbb{E}_{(\mathbf{x},y)\sim P_T} \left[h(\mathbf{x}) \neq y\right]$ **target** domain error

Domain Adaptation objective

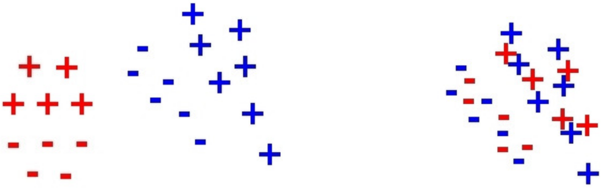
- $h \in \mathcal{H}$ with a **low** $\mathrm{err}_T(h)$

A Transfer Learning Task: Domain Adaptation

Studied case

- **Source** domain
  $LS = \{(\mathbf{x}_i, y_i)\}_{i=1}^{d_l}$ **Labeled** Source sample drawn i.i.d. from P_S
- **Target** domain
  $TS = \{\mathbf{x}_j\}_{j=1}^{d_t}$ **unlabeled** Target Sample drawn i.i.d. from D_T

If h is learned on **source** domain, how does it perform on **target** domain ?

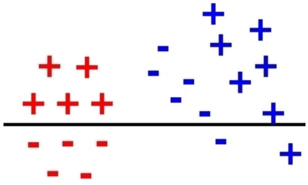


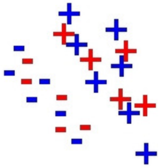
A Transfer Learning Task: Domain Adaptation

Studied case

- **Source** domain
  $LS = \{(\mathbf{x}_i, y_i)\}_{i=1}^{d_l}$ **Labeled** Source sample drawn i.i.d. from P_S
- **Target** domain
  $TS = \{\mathbf{x}_j\}_{j=1}^{d_t}$ **unlabeled** Target Sample drawn i.i.d. from D_T

If h is learned on **source** domain, how does it perform on **target** domain ?



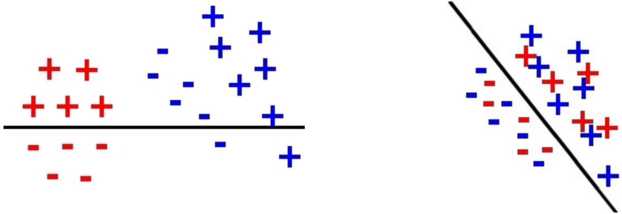


A Transfer Learning Task: Domain Adaptation

Studied case

- **Source** domain
  $LS = \{(\mathbf{x}_i, y_i)\}_{i=1}^{d_l}$ **Labeled** Source sample drawn i.i.d. from P_S
- **Target** domain
  $TS = \{\mathbf{x}_j\}_{j=1}^{d_t}$ **unlabeled** Target Sample drawn i.i.d. from D_T

If h is learned on **source** domain, how does it perform on **target** domain ?



Introduction and Motivation

A Transfer Learning Task: Domain Adaptation (DA)

Learning with Good Similarity Functions (SF)

Modifying the Projection Space for Domain Adaptation

Experimentations

Extended Work: A Little Bit of Theory

Conclusion and Perspectives

A Transfer Learning Task: Domain Adaptation

S. Ben-David *et al.* results

Theorem [Ben-David et al., 2010]

Let $\mathcal{H}$ be an hypothesis space. If D_S and D_T are respectively the marginal distributions of source and target instances, then for all $\delta \in]0,1]$, with probability at least $1-\delta$:

$$\forall h \in \mathcal{H}, \quad \mathrm{err}_T(h) \;\leq\; \mathrm{err}_S(h) \;+\; \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) \;+\; \nu$$

Theorem [Ben-David et al., 2010]

Let $\mathcal{H}$ be an hypothesis space. If D_S and D_T are respectively the marginal distributions of source and target instances, then for all $\delta \in]0,1]$, with probability at least $1-\delta$:

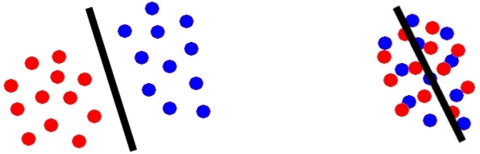
$$\forall h \in \mathcal{H}, \quad \mathrm{err}_T(h) \leq \mathrm{err}_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \nu$$

Theorem [Ben-David et al., 2010]

Let $\mathcal{H}$ be an hypothesis space. If D_S and D_T are respectively the marginal distributions of source and target instances, then for all $\delta \in]0,1]$, with probability at least $1-\delta$:

$$\forall h \in \mathcal{H}, \quad \mathrm{err}_T(h) \ \leq \ \mathrm{err}_S(h) \ + \ \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) \ + \ \nu$$

What is $d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T)$ the $\mathcal{H}\Delta\mathcal{H}$-distance ?
Intuitively



A Transfer Learning Task: Domain Adaptation

S. Ben-David *et al.* results

Theorem [Ben-David et al., 2010]

Let $\mathcal{H}$ be an hypothesis space. If D_S and D_T are respectively the marginal distributions of source and target instances, then for all $\delta \in]0,1]$, with probability at least $1-\delta$:

$$\forall h \in \mathcal{H}, \quad \text{err}_T(h) \leq \text{err}_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \nu$$

What is ν ?

- $\nu = \inf_{h \in \mathcal{H}} \left(\text{err}_S(h) + \text{err}_T(h) \right)$ error of the joint optimal classifier

Theorem [Ben-David et al., 2010]

Let $\mathcal{H}$ be an hypothesis space. If D_S and D_T are respectively the marginal distributions of source and target instances, then for all $\delta \in]0,1]$, with probability at least $1-\delta$:

$$\forall h \in \mathcal{H}, \quad \mathrm{err}_T(h) \ \leq \ \mathrm{err}_S(h) \ + \ \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) \ + \ \nu$$

Idea : Minimizing the bound for building a new projection space

$\Rightarrow$ Explicit projection space defined by a good similarity function

$\Rightarrow$ $\mathcal{H}$ hypothesis space of good similarity functions based classifiers

Outline

A Transfer Learning Task: Domain Adaptation (DA)

Learning with Good Similarity Functions (SF)

Modifying the Projection Space for Domain Adaptation
- An Heuristic Normalization
- An Additional Regularization Term

Experimentations

Extended Work: A Little Bit of Theory

Conclusion and Perspectives

Learning with Good Similarity Functions

(ϵ, γ, τ)-Good Similarity Functions [Balcan et al., 2008a, Balcan et al., 2008b]

$K : X \times X \to [-1;1]$ is an (ϵ, γ, τ)-**good similarity function** for a binary classification problem P if

(i) A $1 - \epsilon$ probability mass of examples $(\mathbf{x}, y)$ satisfy

$$\mathbb{E}_{(\mathbf{x}', y') \sim P}\left[yy'K(\mathbf{x}, \mathbf{x}')|R(\mathbf{x}')\right] \;\geq\; \gamma$$

(ii) $Pr_{\mathbf{x}'}[R(\mathbf{x}')] \geq \tau$
(*Notation:* R set of **reasonable points**)

Intuitively
For a point $(\mathbf{x}_1, y_1) \sim P$, then **on average** for $(\mathbf{x}'_2, y'_2) \in R$

if $y_1 = y'_2$	if $y_1 \neq y'_2$
$\mathbf{x}_1$ is similar to $\mathbf{x}_2$	$\mathbf{x}_1$ is dissimilar to $\mathbf{x}_2$
$K(\mathbf{x}_1, \mathbf{x}'_2) \geq \gamma$	$K(\mathbf{x}_1, \mathbf{x}'_2) \leq -\gamma$

Learning with Good Similarity Functions

Properties [Balcan et al., 2008a, Balcan et al., 2008b]

- $R = \{\mathbf{x}'_j\}_{j=1}^{d_u}$ defines an explicit projection space

$$\phi^R : \begin{cases} X & \rightarrow & \mathbb{R}^{d_u} \\ x & \mapsto & \langle K(\mathbf{x}, \mathbf{x}'_1), \ldots, K(\mathbf{x}, \mathbf{x}'_{d_u}) \rangle \end{cases}$$

- h is learned in this space such as

$$h(\mathbf{x}) = \text{sign}\left[\sum_{j=1}^{d_u} \alpha_j K(\mathbf{x}, \mathbf{x}'_j)\right]$$

  - by solving a linear program
  - with good generalization guarantees

- Generalization of kernels
  - K may be **not** symmetric and **not** positive semi-definite

Outline

A Transfer Learning Task: Domain Adaptation (DA)

Learning with Good Similarity Functions (SF)

Modifying the Projection Space for Domain Adaptation
- An Heuristic Normalization
- An Additional Regularization Term

Experimentations

Extended Work: A Little Bit of Theory

Conclusion and Perspectives

Outline

A Transfer Learning Task: Domain Adaptation (DA)

Learning with Good Similarity Functions (SF)

Modifying the Projection Space for Domain Adaptation
- An Heuristic Normalization
- An Additional Regularization Term

Experimentations

Extended Work: A Little Bit of Theory

Conclusion and Perspectives

Modifying the Projection Space for Domain Adaptation

An Heuristic Normalization of a Similarity Function

Introduction and Motivation
A Transfer Learning Task: Domain Adaptation (DA)
Learning with Good Similarity Functions (SF)
Modifying the Projection Space for Domain Adaptation
Experimentations
Extended Work: A Little Bit of Theory
Conclusion and Perspectives

Recall:

- For a DA task, we want to be performing on the target domain TS
- A SF K must be good on LS relatively to the reasonable points R

Idea: "Insert" target information in K

$\Rightarrow$ Build a new K_N by normalizing a given K relatively to $LS \cup TS$

Heuristic: Given $N = LS \cup TS$, for each $\mathbf{x}' \in R$, K_N must have a mean 0 and a standard deviation 1 on $\mathbf{x} \in N$, *i.e.* K_N is defined by:

$$\forall \mathbf{x}'_j \in R,\ K_N(.,\mathbf{x}'_j) = \begin{cases} \dfrac{K(.,\mathbf{x}'_j) - \mu_{\mathbf{x}'_j}}{\sigma_{\mathbf{x}'_j}} & \text{if} \quad -1 \leq \dfrac{K(.,\mathbf{x}'_j) - \widehat{\mu}_{\mathbf{x}'_j}}{\widehat{\sigma}_{\mathbf{x}'_j}} \leq 1, \\ -1 & \text{if} \quad \dfrac{K(.,\mathbf{x}'_j) - \widehat{\mu}_{\mathbf{x}'_j}}{\widehat{\sigma}_{\mathbf{x}'_j}} \leq -1, \\ 1 & \text{if} \quad 1 \leq \dfrac{K(.,\mathbf{x}'_j) - \widehat{\mu}_{\mathbf{x}'_j}}{\widehat{\sigma}_{\mathbf{x}'_j}}, \end{cases}$$

Outline

A Transfer Learning Task: Domain Adaptation (DA)

Learning with Good Similarity Functions (SF)

Modifying the Projection Space for Domain Adaptation
- An Heuristic Normalization
- An Additional Regularization Term

Experimentations

Extended Work: A Little Bit of Theory

Conclusion and Perspectives

Modifying the Projection Space for Domain Adaptation

An Additional Regularization Term For Moving Closer the Two Distributions

Recall: Minimizing $\mathrm{err}_T(h)$ with the help of the bound

$$\mathrm{err}_T(h) \;\leq\; \mathrm{err}_S(h) \;+\; \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) \;+\; \nu$$

Modifying the Projection Space for Domain Adaptation

An Additional Regularization Term For Moving Closer the Two Distributions

Recall: Minimizing $\mathrm{err}_T(h)$ with the help of the bound

$$\mathrm{err}_T(h) \ \leq \ \mathrm{err}_S(h) \ + \ \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) \ + \ \nu$$

Minimizing $\mathrm{err}_S(h)$ via Balcan optimisation problem for SF with $LS = \{(\mathbf{x}_i, y_i)\}_{i=1}^{d_l}$ and $R' = \{\mathbf{x}'_j\}_{j=1}^{d'_u}$

$$\min_{\alpha_1,\ldots,\alpha_{d_u}} \sum_{i=1}^{d_l} \left[1 - y_i \sum_{j=1}^{d'_u} \alpha_j K(\mathbf{x}_i, \mathbf{x}'_j) \right]_+ \ + \ \lambda \|\boldsymbol{\alpha}\|_1$$

$[1-a]_+ = \max(1-a; 0)$ is the hinge loss
$\|\boldsymbol{\alpha}\|_1 = \sum_{j=1}^{d'_u} |\alpha_j|$

$\mathrm{err}_T(h) \ \leq \ \mathrm{err}_S(h) \ + \ \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) \ + \ \nu$

Modifying the Projection Space for Domain Adaptation

An Additional Regularization Term For Moving Closer the Two Distributions

Minimizing $d_{\mathcal{H}\Delta\mathcal{H}}(U_S, U_T)$ ($U_S \sim D_S$ and $U_T \sim D_T$)

$\Rightarrow \mathcal{C}_{ST}$ a pair set $(\mathbf{x}_s, \mathbf{x}_t) \in U_S \times U_T$

Building a new projection $\phi^{R'}_{new}$

s.t. $\mathbf{x}_s$ and $\mathbf{x}_t$ be not separable

s.t. with the result that $|h(\mathbf{x}_s) - h(\mathbf{x}_t)| \approx 0$

$$\left|\sum_{j=1}^{d'_u} \alpha_j K(\mathbf{x}_s,\mathbf{x}'_j) - \sum_{j=1}^{d'_u} \alpha_j K(\mathbf{x}_t,\mathbf{x}'_j)\right| \leq \left\|({}^t\phi^{R'}(\mathbf{x}_s) - {}^t\phi^{R'}(\mathbf{x}_t))\mathrm{diag}(\boldsymbol{\alpha})\right\|_1$$

$$\mathrm{err}_T(h) \leq \mathrm{err}_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \nu$$

Modifying the Projection Space for Domain Adaptation

An Additional Regularization Term For Moving Closer the Two Distributions

Minimizing $d_{\mathcal{H}\Delta\mathcal{H}}(U_S, U_T)$ ($U_S \sim D_S$ and $U_T \sim D_T$)

$\Rightarrow \mathcal{C}_{ST}$ a pair set $(\mathbf{x}_s, \mathbf{x}_t) \in U_S \times U_T$

Building a new projection $\phi^{R'}_{new}$

s.t. $\mathbf{x}_s$ and $\mathbf{x}_t$ be not separable

s.t. with the result that $|h(\mathbf{x}_s) - h(\mathbf{x}_t)| \approx 0$

$$\left|\sum_{j=1}^{d'_u} \alpha_j K(\mathbf{x}_s,\mathbf{x}'_j) - \sum_{j=1}^{d'_u} \alpha_j K(\mathbf{x}_t,\mathbf{x}'_j)\right| \leq \underbrace{\left\|({}^t\phi^{R'}(\mathbf{x}_s) - {}^t\phi^{R'}(\mathbf{x}_t))\mathrm{diag}(\boldsymbol{\alpha})\right\|_1}_{\left\|{}^t\phi^{R'}_{new}(\mathbf{x}_s) - {}^t\phi^{R'}_{new}(\mathbf{x}_t)\right\|_1}$$

$$\Rightarrow \phi^{R'}_{new}(.) = \langle \underbrace{\alpha_1 K(., x'_1)}_{K_{new}(., x'_1)}, \ldots, \underbrace{\alpha_{d_u} K(., x'_{d_u})}_{K_{new}(., x'_{d_u})} \rangle$$

$\mathrm{err}_T(h) \leq \mathrm{err}_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \nu$

Modifying the Projection Space for Domain Adaptation

An Additional Regularization Term For Moving Closer the Two Distributions

Minimizing $d_{\mathcal{H}\Delta\mathcal{H}}(U_S, U_T)$ ($U_S \sim D_S$ and $U_T \sim D_T$)

$\Rightarrow$ $\mathcal{C}_{ST}$ a pair set $(\mathbf{x}_s, \mathbf{x}_t) \in U_S \times U_T$

Building a new projection $\phi^{R'}_{new}$

s.t. $\mathbf{x}_s$ and $\mathbf{x}_t$ be not separable

s.t. with the result that $|h(\mathbf{x}_s) - h(\mathbf{x}_t)| \approx 0$

$$\left|\sum_{j=1}^{d'_u} \alpha_j K(\mathbf{x}_s,\mathbf{x}'_j) - \sum_{j=1}^{d'_u} \alpha_j K(\mathbf{x}_t,\mathbf{x}'_j)\right| \leq \underbrace{\left\|\left({}^t\phi^{R'}(\mathbf{x}_s) - {}^t\phi^{R'}(\mathbf{x}_t)\right)\mathrm{diag}(\boldsymbol{\alpha})\right\|_1}_{\left\|{}^t\phi^{R'}_{new}(\mathbf{x}_s) - {}^t\phi^{R'}_{new}(\mathbf{x}_t)\right\|_1}$$

$$\Rightarrow \phi^{R'}_{new}(.) = \langle \underbrace{\alpha_1 K(., x'_1)}_{K_{new}(., x'_1)}, \ \ldots \ , \underbrace{\alpha_{d_u} K(., x'_{d_u})}_{K_{new}(., x'_{d_u})} \rangle$$

$\Rightarrow$ New regularization term

$\mathrm{err}_T(h) \leq \mathrm{err}_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \nu$

With $LS = \{(\mathbf{x}_i, y_i)\}_{i=1}^{d_l}$ (i.i.d. from P_S) and $R' = \{\mathbf{x}'_j\}_{j=1}^{d'_u}$

Building the $\phi_{new}^{R'}$ space with the help of $\boldsymbol{\alpha}$ infered by

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{d_l} \left[1 - y_i \sum_{j=1}^{d'_u} \alpha_j K(\mathbf{x}_i, \mathbf{x}'_j)\right]_+ + \lambda \|\boldsymbol{\alpha}\|_1 + \beta \sum_{(\mathbf{x}_s, \mathbf{x}_t) \in \mathcal{C}_{ST}} \|({}^t\phi^{R'}(\mathbf{x}_s) - {}^t\phi^{R'}(\mathbf{x}_t))\,\mathrm{diag}(\boldsymbol{\alpha})\|_1$$

$$\mathrm{err}_T(h) \leq \mathrm{err}_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \nu$$

Introduction and Motivation
A Transfer Learning Task: Domain Adaptation (DA)
Learning with Good Similarity Functions (SF)
Modifying the Projection Space for Domain Adaptation
Experimentations
Extended Work: A Little Bit of Theory
Conclusion and Perspectives

Modifying the Projection Space for Domain Adaptation

An Additional Regularization Term For Moving Closer the Two Distributions

With $LS = \{(\mathbf{x}_i, y_i)\}_{i=1}^{d_l}$ (i.i.d. from P_S) and $R' = \{\mathbf{x}'_j\}_{j=1}^{d'_u}$

Building the $\phi_{new}^{R'}$ space with the help of $\boldsymbol{\alpha}$ infered by

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{d_l} \left[1 - y_i \sum_{j=1}^{d'_u} \alpha_j K(\mathbf{x}_i, \mathbf{x}'_j) \right]_+ + \lambda \|\boldsymbol{\alpha}\|_1$$
$$+ \beta \sum_{(\mathbf{x}_s, \mathbf{x}_t) \in \mathcal{C}_{ST}} \|({}^t\phi^{R'}(\mathbf{x}_s) - {}^t\phi^{R'}(\mathbf{x}_t)) \, \mathrm{diag}(\boldsymbol{\alpha})\|_1$$

- Validation of hyperparameters, of reweighting, of $\mathcal{C}_{ST}$?

$\mathrm{err}_T(h) \ \leq \ \mathrm{err}_S(h) \ + \ \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) \ + \ \nu$

Modifying the Projection Space for Domain Adaptation

Reverse Classifier h^r and Validation

Problem: No label on target domain

Solution: Kind of "reverse" validation [Zhong et al., 2010]
With the reverse classifier h^r

Introduction and Motivation

A Transfer Learning Task: Domain Adaptation (DA)

Learning with Good Similarity Functions (SF)

Modifying the Projection Space for Domain Adaptation

Experimentations

Extended Work: A Little Bit of Theory

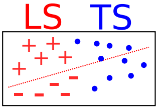
Conclusion and Perspectives

Modifying the Projection Space for Domain Adaptation

Reverse Classifier h^r and Validation

Problem: No label on target domain

Solution: Kind of "reverse" validation [Zhong et al., 2010]
With the reverse classifier h^r

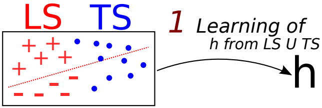


Modifying the Projection Space for Domain Adaptation

Reverse Classifier h^r and Validation

Problem: No label on target domain

Solution: Kind of "reverse" validation [Zhong et al., 2010]
With the reverse classifier h^r

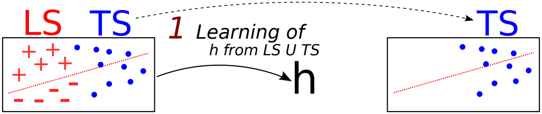


Modifying the Projection Space for Domain Adaptation

Reverse Classifier h^r and Validation

Problem: No label on target domain

Solution: Kind of "reverse" validation [Zhong et al., 2010]
With the reverse classifier h^r

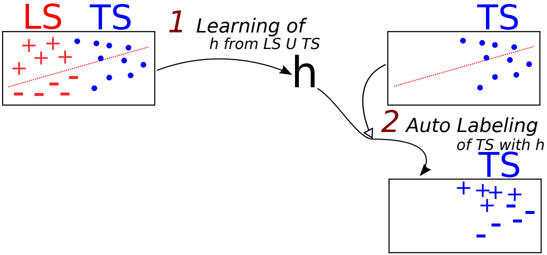


Modifying the Projection Space for Domain Adaptation

Reverse Classifier h^r and Validation

Problem: No label on target domain

Solution: Kind of "reverse" validation [Zhong et al., 2010]
With the reverse classifier h^r



Introduction and Motivation

A Transfer Learning Task: Domain Adaptation (DA)

Learning with Good Similarity Functions (SF)

Modifying the Projection Space for Domain Adaptation

Experimentations

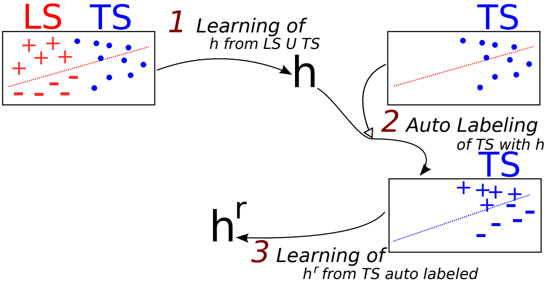
Extended Work: A Little Bit of Theory

Conclusion and Perspectives

Modifying the Projection Space for Domain Adaptation
Reverse Classifier h^r and Validation

Problem: No label on target domain

Solution: Kind of "reverse" validation [Zhong et al., 2010]
With the reverse classifier h^r

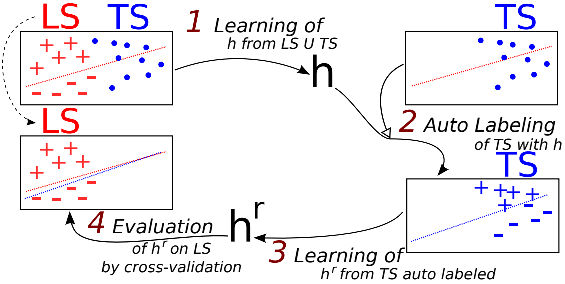


Modifying the Projection Space for Domain Adaptation

Reverse Classifier h^r and Validation

Problem: No label on target domain

Solution: Kind of "reverse" validation [Zhong et al., 2010]
With the reverse classifier h^r



- Two domains are related $\Rightarrow$ h^r performs well on the source domain

[Bruzzone and Marconcini, 2010]

Outline

A Transfer Learning Task: Domain Adaptation (DA)

Learning with Good Similarity Functions (SF)

Modifying the Projection Space for Domain Adaptation
- An Heuristic Normalization
- An Additional Regularization Term

Experimentations

Extended Work: A Little Bit of Theory

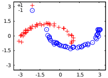
Conclusion and Perspectives

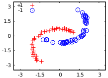
Experimentations

Experimental Setup

- Similarity function:
  - K Gaussian kernel
  - K_{ST} Normalization of K according to $LS \cup TS$
- Comparison of performances of K and K_{ST}
  - with the new regularization and without

1. Toy problem "inter-twinning moons"





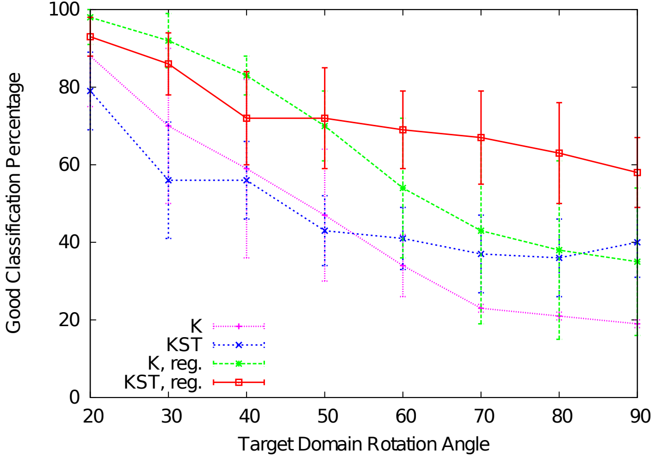
   - 1 source domain
   - 8 different target domains according to 8 rotation angles
   - 10 draws for each angle
   - Performances on a test set of 1500 target instances

2. Image annotation

   - Source domain: PascalVOC 2007
   - Target domain: TrecVid 2007
   - F-measures on target domain

Experimentations

Inter-twinning moons: results



Introduction and Motivation

A Transfer Learning Task: Domain Adaptation (DA)

Learning with Good Similarity Functions (SF)

Modifying the Projection Space for Domain Adaptation

Experimentations

Extended Work: A Little Bit of Theory

Conclusion and Perspectives

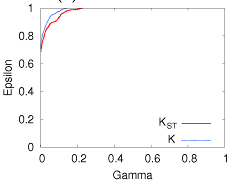
Experimentations

Inter-twinning moons: estimation of the similarity function goodness on *TS*

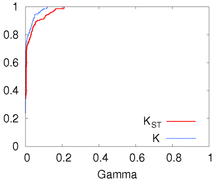
ϵ on *TS* as a function of γ
i.e. for a given γ, ϵ is the proportion of $\mathbf{x} \in TS$ s.t.:

$$\sum_{\mathbf{x}'_j \in R'} y_i y'_j K(\mathbf{x}_i, \mathbf{x}'_j) < \gamma$$

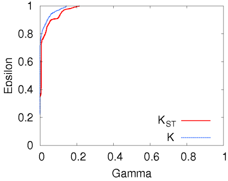
(a) For a 20° task.



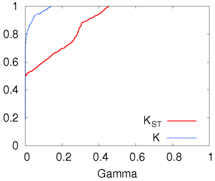
(b) For a 30° task.



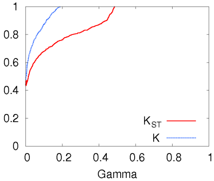
(c) For a 40° task.



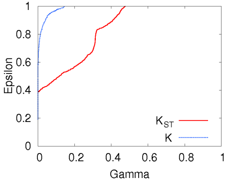
(d) For a 50° task.



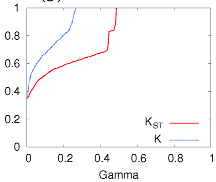
(e) For a 60° task.



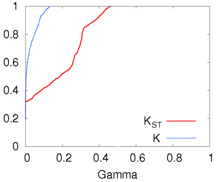
(f) For a 70° task.



(g) For a 80° task.



(h) For a 90° task.



Introduction and Motivation

A Transfer Learning Task: Domain Adaptation (DA)

Learning with Good Similarity Functions (SF)

Modifying the Projection Space for Domain Adaptation

Experimentations

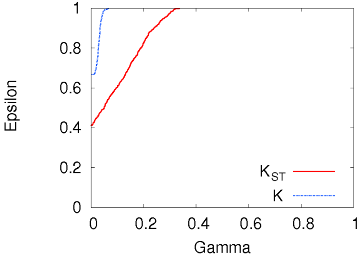
Extended Work: A Little Bit of Theory

Conclusion and Perspectives

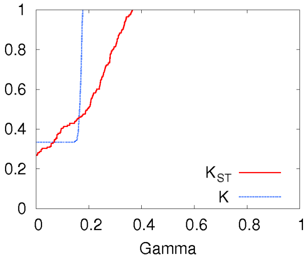
Experimentations

Images corpus: results

CONC.	BOAT	BUS	CAR	MONITOR	PERSON	PLANE	AVG.
SF without distance regularization							
K	0.0279	0.1806	0.5214	0.2477	0.4971	0.5522	0.3378
K_{ST}	0.4731	0.4632	0.5316	0.3664	0.3776	0.5635	0.4626
SF with distance regularization							
K	0.2006	0.1739	0.5125	0.2744	0.5037	0.5192	0.3640
K_{ST}	**0.4857**	**0.4891**	**0.5452**	**0.3989**	**0.5353**	**0.6375**	**0.5153**



(a) For the concept BOAT



(b) For the concept PLANE.

Outline

A Transfer Learning Task: Domain Adaptation (DA)

Learning with Good Similarity Functions (SF)

Modifying the Projection Space for Domain Adaptation
- An Heuristic Normalization
- An Additional Regularization Term

Experimentations

Extended Work: A Little Bit of Theory

Conclusion and Perspectives

Extended Work: A little Bit of Theory

Sparsity Analysis

Recall:

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{d_l}\left[1-y_i\sum_{j=1}^{d_u'}\alpha_j K(\mathbf{x}_i,\mathbf{x}_j')\right]_+ +\lambda\|\boldsymbol{\alpha}\|_1+\beta\sum_{(\mathbf{x}_s,\mathbf{x}_t)\in\mathcal{C}_{ST}}\|({}^t\phi^{R'}(\mathbf{x}_s)-{}^t\phi^{R'}(\mathbf{x}_t))\,\mathrm{diag}(\boldsymbol{\alpha})\|_1$$

Lemma

Let $B_R = \min_{\mathbf{x}_j'\in R}\left\{\max_{(\mathbf{x}_s,\mathbf{x}_t)\in\mathcal{C}_{ST}} |K(\mathbf{x}_s,\mathbf{x}_j') - K(\mathbf{x}_t,\mathbf{x}_j')|\right\} > 0.$

If $\boldsymbol{\alpha}^*$ is the optimal solution of our problem, then,

$$\|\boldsymbol{\alpha}^*\|_1 \leq \frac{1}{\beta B_R + \lambda}.$$

$\Rightarrow$ The sparsity depends on the hyperparameters and B_R

$\Rightarrow$ The domains are far $\Rightarrow$ The difference between coordinates is high

- $\rightarrow$ B_R tends to be high
- $\rightarrow$ increase of the sparsity

Extended Work: A little Bit of Theory

Generalization Bounds

- Investigation of algorithmic robustness [Xu and Mannor, 2010]
  - Idea: *"if a testing sample is similar to a training sample then the testing error is close to the training error"* (in a classical ML setting)

$\Rightarrow$ Our method is robust on the source domain

$\Rightarrow$ Generalization bound:

$$err_T(h) \leq \hat{err}_S(h) + \frac{N_\eta}{\beta B_R + \lambda} + \sqrt{\frac{4M_\eta \ln 2 + 2\ln\frac{1}{\delta}}{d_l}} + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \nu,$$

Outline

A Transfer Learning Task: Domain Adaptation (DA)

Learning with Good Similarity Functions (SF)

Modifying the Projection Space for Domain Adaptation
- An Heuristic Normalization
- An Additional Regularization Term

Experimentations

Extended Work: A Little Bit of Theory

Conclusion and Perspectives

Conclusion and Perspectives

Conclusion

- Domain Adaptation exploiting a similarity based projection space
  1. Normalization of a SF according to the target domain
  2. Addition of a new regularization term for moving closer the domains
     - With a "reverse" validation
     - With generalization guarantees
     - Infers sparse classifiers related to the task difficulty

  Remark Extended work (ICDM'11): an iterative method improves the results

⇒ The SF helps to build a relevant projection space for adaptation

Perspectives

- Influence of target labels
- Design SF for Domain Adaptation
- Other applications

Introduction and Motivation

A Transfer Learning Task: Domain Adaptation (DA)

Learning with Good Similarity Functions (SF)

Modifying the Projection Space for Domain Adaptation

Experimentations

Extended Work: A Little Bit of Theory

Conclusion and Perspectives

Thank you for your attention.

Introduction and Motivation

A Transfer Learning Task: Domain Adaptation (DA)

Learning with Good Similarity Functions (SF)

Modifying the Projection Space for Domain Adaptation

Experimentations

Extended Work: A Little Bit of Theory

Conclusion and Perspectives

References I

Balcan, M.-F., Blum, A. and Srebro, N. (2008a).
Improved Guarantees for Learning via Similarity Functions.
In Proceedings of COLT pp. 287–298,.

Balcan, M.-F., Blum, A. and Srebro, N. (2008b).
A theory of learning with similarity functions.
Machine Learning *72*, 89–112.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F. and Vaughan, J. (2010).
A theory of learning from different domains.
Machine Learning Journal *79*, 151–175.

Bruzzone, L. and Marconcini, M. (2010).
Domain Adaptation Problems: A DASVM Classification Technique and a Circular Validation Strategy.
IEEE Trans. Pattern Anal. Mach. Intell. *32*, 770–787.

Xu, H. and Mannor, S. (2010).
Robustness and Generalization.
In Proceedings of COLT pp. 503–515,.

References II

Zhong, E., Fan, W., Yang, Q., Verscheure, O. and Ren, J. (2010).
Cross Validation Framework to Choose amongst Models and Datasets for Transfer Learning.
In Proceedings of ECML-PKDD (Part III) pp. 547–562, Springer.

Introduction and Motivation

A Transfer Learning Task: Domain Adaptation (DA)

Learning with Good Similarity Functions (SF)

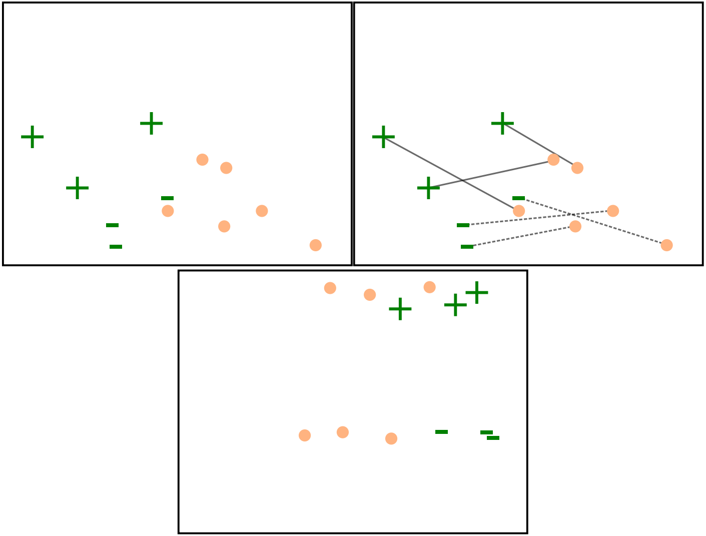
Modifying the Projection Space for Domain Adaptation

Experimentations

Extended Work: A Little Bit of Theory

Conclusion and Perspectives

Appendix
Example



Recall: We solve

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{d_l} \left[1 - y_i \sum_{j=1}^{d'_u} \alpha_j K(\mathbf{x}_i, \mathbf{x}'_j) \right]_+ + \lambda \|\boldsymbol{\alpha}\|_1$$
$$+ \beta \sum_{(\mathbf{x}_s, \mathbf{x}_t) \in \mathcal{C}_{ST}} ({}^t\phi^{R'}(\mathbf{x}_s) - {}^t\phi^{R'}(\mathbf{x}_t)) \,\mathrm{diag}(\boldsymbol{\alpha})\|_1$$

Hypothesis:

- $\forall \mathbf{x}'_j \in R', \max_{(\mathbf{x}_s, \mathbf{x}_t) \in \mathcal{C}_{ST}} |K(\mathbf{x}_s, \mathbf{x}'_j) - K(\mathbf{x}_t, \mathbf{x}'_j)| > 0$
- (X, ρ) is a compact metric space
- K is a continuous similarity function on its first argument

Introduction and Motivation

A Transfer Learning Task: Domain Adaptation (DA)

Learning with Good Similarity Functions (SF)

Modifying the Projection Space for Domain Adaptation

Experimentations

Extended Work: A Little Bit of Theory

Conclusion and Perspectives

Lemma

For any $\lambda > 0$, $\beta > 0$ and any pair set $\mathcal{C}_{ST}$ verifying the previous hypothesis, let $B_R = \min_{\mathbf{x}'_j \in R} \left\{ \max_{(\mathbf{x}_s, \mathbf{x}_t) \in \mathcal{C}_{ST}} |K(\mathbf{x}_s, \mathbf{x}'_j) - K(\mathbf{x}_t, \mathbf{x}'_j)| \right\}$.

If $\boldsymbol{\alpha}^*$ is the optimal solution of our problem, then,

$$\|\boldsymbol{\alpha}^*\|_1 \leq \frac{1}{\beta B_R + \lambda}.$$

$\Rightarrow$ The sparsity depends on the hyperparameters and B_R

$\Rightarrow$ The domains are far $\Rightarrow$ The difference between coordinates is high

$\rightarrow$ B_R tends to be high

$\rightarrow$ increase of the sparsity

- Algorithmic robustness [Xu and Mannor, 2010]
  - Idea: *"if a testing sample is similar to a training sample then the testing error is close to the training error"* (in a classical ML setting)
  - $\mathbf{x}_s \in LS$ and $\mathbf{x}_t \in TS$ are close (according to a metric)

  $\Rightarrow \big|L(h, \mathbf{x}_s) - L(h, \mathbf{x}_t)\big| \leq \epsilon$

$\Rightarrow$ Generalization bounds

- Even if the robustness property is fulfilled for only a subpart of LS

Definition of algorithmic robustness [Xu and Mannor, 2010]

Given a learning sample LS, an algorithm $\mathcal{A}$ is $(\mathbf{M}, \boldsymbol{\epsilon}(\mathbf{LS}))$ **robust** if $X \times Y$ can be partitioned into M disjoint sets, denoted as $\{C_i\}_{i=1}^{M}$, such that $\forall s \in LS$,

$$s, u \in C_i \Rightarrow \ \left| L(h, s) - L(h, u) \right| \leq \epsilon(LS),$$

with h the model learned from LS, L the loss function of $\mathcal{A}$.

Theorem [Xu and Mannor, 2010]

If $LS = \{(\mathbf{x}_i, y_i)\}_{i=1}^{d_l}$ is drawn i.i.d. from a distribution P and if the algorithm $\mathcal{A}$ is $(M,\epsilon(LS))$ robust, then for any $\delta > 0$, with probability at least 1-δ,

$$\mathrm{err}_P(\mathcal{A}_{LS}) \ \leq \widehat{\mathrm{err}}_P(\mathcal{A}_{LS}) + \epsilon(LS) + L^{UP}\sqrt{\frac{2M\ln 2 + 2\ln(1/\delta)}{d_l}},$$

where err_P and $\widehat{\mathrm{err}}_P$ are respectively the expected and the empirical errors over P, L being upper bounded by L^{UP}.

Introduction and Motivation
A Transfer Learning Task: Domain Adaptation (DA)
Learning with Good Similarity Functions (SF)
Modifying the Projection Space for Domain Adaptation
Experimentations
Extended Work: A Little Bit of Theory
Conclusion and Perspectives

Theorem

If $LS = \{(\mathbf{x}_i, y_i)\}_{i=1}^{d_l}$ is drawn i.i.d. from P_S, then our method is $(\mathbf{2M}_\eta, \frac{\mathbf{N}_\eta}{\beta \mathbf{B_R} + \lambda})$ **robust on the source domain $\mathbf{P_S}$**, where $N_\eta = \max_{\substack{\mathbf{x}_a, \mathbf{x}_b \sim D_S \\ \rho(\mathbf{x}_a, \mathbf{x}_b) \leq \eta}} \|{}^t\phi^R(\mathbf{x}_a) - {}^t\phi^R(\mathbf{x}_b)\|_\infty$ with $\eta > 0$ and M_η is the η-covering number of X. Thus for every h in the hypothesis class $\mathcal{H}$ of SF classifiers, for any $\delta > 0$, with probability at least $1 - \delta$,

$$err_S(h) \leq \hat{err}_S(h) + \frac{N_\eta}{\beta B_R + \lambda} + \sqrt{\frac{4M_\eta \ln 2 + 2 \ln \frac{1}{\delta}}{d_l}}.$$

Thus,

$$err_T(h) \leq \hat{err}_S(h) + \frac{N_\eta}{\beta B_R + \lambda} + \sqrt{\frac{4M_\eta \ln 2 + 2 \ln \frac{1}{\delta}}{d_l}} + d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \nu,$$

where ν is the joint error over the domains, $d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T)$ is the $\mathcal{H}\Delta\mathcal{H}$-distance between the marginal distributions.

Appendix

A Way to Lighten the Search of the Projection Space

Recall: $\phi^{R'}_{init}(.) \;=\; \langle\; K_{init}(., x'_1)\;,\;\ldots\;,\; K_{init}(., x'_{d_u})\;\rangle$

$\hookrightarrow$ Learning $\boldsymbol{\alpha}$ thanks to the regularization term

$$\sum_{(\mathbf{x}_s,\mathbf{x}_t)\in\mathcal{C}_{ST}} \underbrace{\left\|\left({}^t\phi^{R'}_{init}(\mathbf{x}_s) - {}^t\phi^{R'}_{init}(\mathbf{x}_t)\right)\mathrm{diag}(\boldsymbol{\alpha})\right\|_1}_{\left\|{}^t\phi^{R'}_{new}(\mathbf{x}_s)-{}^t\phi^{R'}_{new}(\mathbf{x}_t)\right\|_1}$$

$$\Rightarrow \phi^{R'}_{new}(.) \;=\; \langle\; \underbrace{\alpha_1 K_{init}(., x'_1)}_{K_{new}(., x'_1)}\;,\;\ldots\;,\; \underbrace{\alpha_{d_u} K_{init}(., x'_{d_u})}_{K_{new}(., x'_{d_u})}\;\rangle$$

Problem: Testing all the possible pair set $\mathcal{C}_{ST}$ is clearly intractable.

Solution: We iterate the learning process in the new $\phi^{R'}_{new}$-space

$\hookrightarrow$ Stopping criterion ?

Appendix

A Way to Lighten the Search of the Projection Space

Recall : DA Bound: $\mathrm{err}_T(h) \ \leq \ \mathrm{err}_S(h) \ + \ \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) \ + \ \nu$

Joint error $\nu = \inf_{h\in\mathcal{H}} \left(\mathrm{err}_S(h) + \mathrm{err}_T(h)\right) \quad \Leftrightarrow \quad$ Adaptation ability

Recall : DA Bound: $\text{err}_T(h) \ \leq \ \text{err}_S(h) \ + \ \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) \ + \ \nu$

Joint error $\nu = \inf_{h\in\mathcal{H}} \left(\text{err}_S(h) + \text{err}_T(h) \right)$ $\Leftrightarrow$ Adaptation ability
$\Rightarrow$ Stopping criterion

Problem: No label on the target domain

Solution: At each iteration l, we empirically estimate ν

$$\widehat{\nu}_l = \widehat{\text{err}}_S(h_l^r) + \widehat{\text{err}}_T(h_l^r)$$

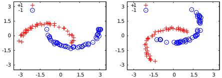
where $\widehat{\text{err}}_T(h_l^r)$ is the error of h_l^r on TS auto-labeled by h_l
We select parameters associated with the minimal $\widehat{\nu}_l$

$\Rightarrow$ Stop at iteration l, if $\widehat{\nu}_{l+1}$ increases or converges comparing to $\widehat{\nu}_l$
Return $h_l(.)$ with the minimal $\widehat{\nu}_l$

Appendix
Experimental Setup

- Similarity function:
  - K Gaussian kernel
  - K_{ST} Normalization of K according to $LS \cup TS$
- Comparison with SVM, TSVM, DASVM and SF
  - Performances and model sizes

1. Toy problem "inter-twinning moons"



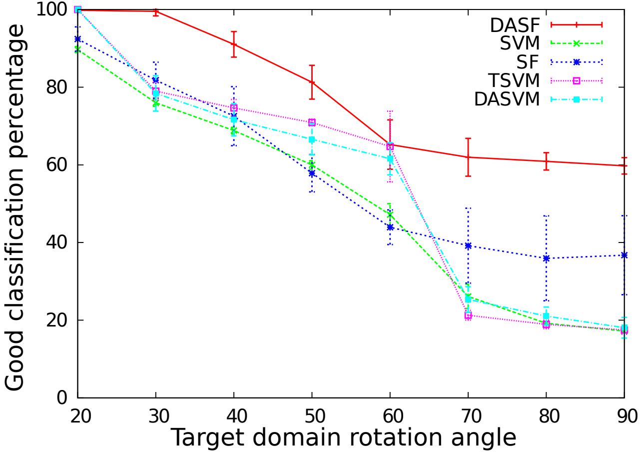
   - 1 source domain
   - 8 different target domains according to 8 rotation angles
   - 10 draws for each angle
   - Performances on a test set of 1500 target instances

2. Image annotation

   - Source domain: PascalVOC 2007 with ratio $+/-$ de 1/3
   - Two target domains:
     - Different ratio $+/-$: PascalVOC 2007 Test
     - Same ratio $+/-$: TrecVid 2007
   - F-measures on target domain

Appendix

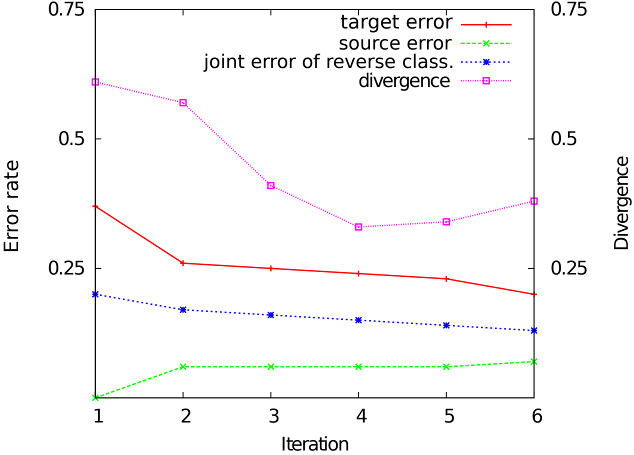
Inter-twinning moons: results



Size	10	10	9	8	4	4	4	3

Appendix

Inter-twinning moons: an execution example

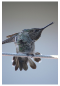


Appendix

Images corpus: results

VOC vs VOC: Reasonable points for the concept PERSON



Results

		SVM	SF	TSVM	DASVM	DASF
VOC vs VOC						
Avg. on 20 conc.	F-meas.	0.22	0.19	0.17	0.20	**0.25**
	Size	642	210	705	622	**200**
VOC vs Trec						
BOAT	F-meas	0.56	0.49	0.56	0.52	**0.57**
	Size	351	214	498	202	**120**
CAR	F-meas.	0.43	0.50	0.52	**0.55**	**0.55**
	Size	1096	**176**	631	627	254
MONITOR	F-meas.	0.19	0.34	0.37	0.30	**0.42**
	Size	698	246	741	523	**151**
PERSON	F-meas.	0.52	0.45	0.46	0.54	**0.57**
	Size	951	226	1024	274	**19**
PLANE	F-meas.	0.32	0.54	0.61	0.52	**0.66**
	Size	428	178	259	450	**7**
Avg. on the 5 conc.	F-meas.	0.40	0.47	0.50	0.49	**0.55**
	Size	705	208	631	415	**110**

Introduction and Motivation

A Transfer Learning Task: Domain Adaptation (DA)

Learning with Good Similarity Functions (SF)

Modifying the Projection Space for Domain Adaptation

Experimentations

Extended Work: A Little Bit of Theory

Conclusion and Perspectives