JDCOT : an Algorithm for Transfer Learning in Incomparable Domains using Optimal Transport

Marion Jeamart *, Renan Bernard *, Nicolas Courty *,
Chloé Friguet * & **Valérie Garès****

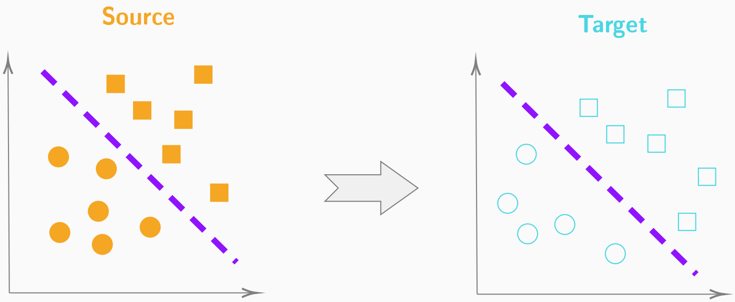
Journée MAS 2022 - Rouen
30/08/2022

* *Univ. Bretagne-Sud, UMR 6074, IRISA, Vannes*
** *INSA, UMR 6625, IRMAR, Rennes*

Introduction

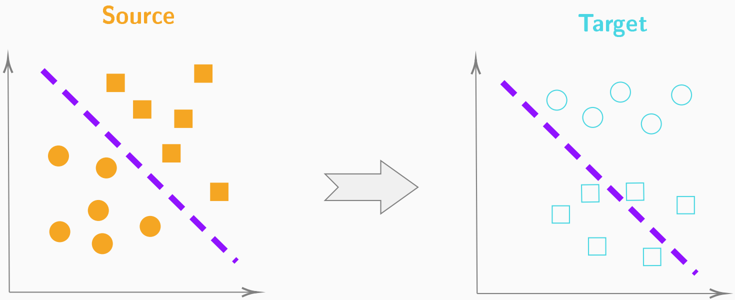
Introduction

- Usual learning process
  - learn a model on source data $(X^S, Y^S) \in \mathbb{R}^{n^S \times d^S} \times \mathcal{C}$
  - use the model on target data $(X^T, Y^T) \in \mathbb{R}^{n^T \times d^T} \times \mathcal{C}$



Introduction

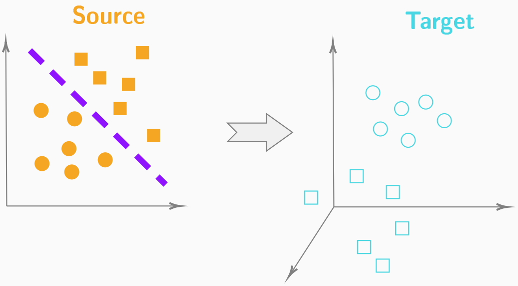
- Usual learning process
  - learn a model on source data $(X^S, Y^S) \in \mathbb{R}^{n^S \times d^S} \times \mathcal{C}$
  - use the model on target data $(X^T, Y^T) \in \mathbb{R}^{n^T \times d^T} \times \mathcal{C}$



✘ If source and target data do not have the same distribution?

Domain Adaptation (DA)

- **Transfer** learnt knowledge from source domain to target domain : same task (classification), different (but related) domains
  - trained model becomes more robust when being used on data lying in another domain
  - less labelled data needed in target domain
- **Heterogeneous** domain adaptation (HDA) : source and target domains are represented by **different features spaces**



In DA literature

- Strategies
  - Project both data into a common subspace by jointly learning the common subspace and a classifier
  - Jointly perform implicit data reconstruction and learn a classifier
- Supervision settings

	Y^S	Y^T
Unsupervised DA	observed	unobserved
Semi-supervised DA	observed	partially observed
Partial DA	partially observed	partially observed

Our proposal
Deal with **heterogeneous domain adaptation** using **optimal transport**

OT for DA

Optimal transport (OT)

- Optimisation method (Peyré and Cuturi, 2018)
  - Distance between two probability measures (Wasserstein distance)
  - Loss in many optimisation problems and approximation algorithms
- Kantorovich formulation to find a coupling matrix γ between
  - $X^S = \{(x_i^S, w_i^S)_{i=1\dots n^S}, \sum_{i=1}^{n^S} w_i^S = 1\}$
  - $X^T = \{(x_j^T, w_j^T)_{j=1\dots n^T}, \sum_{j=1}^{n^T} w_j^T = 1\}$

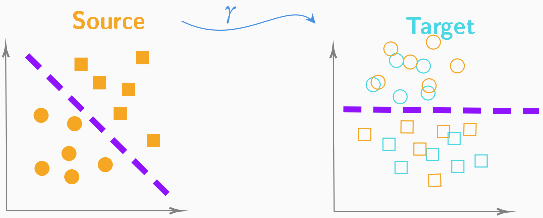
$$\gamma = OT(w^S, w^T, C) = \underset{P \in U(w^S, w^T)}{\operatorname{argmin}} \sum_{i,j} C_{ij} P_{ij}$$

$U(w^S, w^T)$: set of matrices $P \in \mathbb{R}_+^{n^S \times n^T}$ so that $\sum_{i=1}^{n^S} P_{ij} = w_j^T$, $\forall j = 1 \dots n^T$ and $\sum_{j=1}^{n^T} P_{ij} = w_i^S$, $\forall i = 1 \dots n^S$

C: a cost matrix

OT for domain adaptation (Courty et al., 2016)

- **Solve the OT problem** $\gamma = OT\left(1/n^S; 1/n^T; d(X^S; X^T)\right)$
  - Assumption : existence of a transfer map M from source to target domain distributions so that $\mathbb{P}(Y^T|X^T) = \mathbb{P}(Y^S|M(X^S))$ and $\mathbb{P}(X^T) = \mathbb{P}(M(X^S))$
- **Transport source data** onto the target domain (barycentric mapping) with γ
- **Learn the target classifier** with the transported source data



Joint Distribution OT (JDOT) (Courty et al., 2017)

Simultaneous optimisation of the **coupling matrix** γ and the **classifier** f

$$\min_{\gamma,f} \sum_{i,j} \Big[\alpha \ d(x_i^S, x_j^T) + \mathcal{L}(y_i^S, f(x_j^T))\Big] \gamma_{ij}$$

Assumption: existence of a transfer map from source domain joint distribution $\mathbb{P}^S(X;Y)$ into target joint distribution $\mathbb{P}^T(X;Y)$

Algorithm 1: Block Coordinate Descent (BCD) for JDOT

initialization: $Y^T_{pred} = f^T_{init}(X^T)$

for $k=1..itermax$ **do**

Update **transport map**:

$\gamma = OT(w^S, w^T, \alpha d(X^S, X^T) + \mathcal{L}(Y^S, Y^T_{pred}))$

Update target label: *// label propagation*

$\hat{Y}^T = n_T \gamma Y^S$

train **classifier** f with $(X^T, \hat{Y}^T)$

predict $Y^T_{pred} = f(X^T)$

✘ Does not address the *heterogeneous* domain adaptation problem

OT for HDA

Co-OT for heterogeneous domains (Redko et al., 2020)

- Simultaneous solving of the OT problem on the samples (γ^s) and on the variables (γ^v)

$$\min_{\gamma^s,\gamma^v} \sum_{i,j,k,\ell} d(x^S_{i,k}, x^T_{j,\ell}) {\gamma^s}_{i,j} {\gamma^v}_{k,\ell}$$

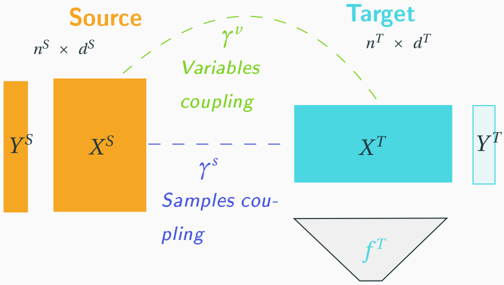
- Use label propagation (with γ^s) to get $\hat{Y}^T$

Our proposal
Use the principle of CoOT to adapt JDOT to the HDA framework

Joint Distribution Co-Optimal Transport (JDCOT)

Simultaneous solving of the OT problem on the samples (γ^s), on the variables (γ^v) and learn the classifier (f^T) in target domain

$$\min_{\gamma^s,\gamma^v,f^T} \sum_{i,j,k,\ell} \left[\alpha \; d({x^S}_{i,k}, {x^T}_{j,\ell}) + \mathcal{L}\Big(y_i^S, f^T(x_j^T)\Big) \right] {\gamma^s}_{ij} {\gamma^v}_{k\ell}$$



Optimisation : Block Coordinate Descent

Algorithm 2: BCD for JDCOT.

Initialisation : $\gamma^s = \gamma^s_{init},\ \gamma^v = \gamma^v_{init},\ Y^T_{pred} = f^T_{init}(X^T)$

for *k=1...itermax* **do**

Update **transport maps**:

$\gamma^s = OT(w^S, w^T, \alpha D(X^S, X^T) \otimes \gamma^v + \mathcal{L}(Y^S, Y^T_{pred}))$

$\gamma^v = OT(w'^S, w'^T, D(X^S, X^T) \otimes \gamma^s)$

Update target label: *// label propagation*

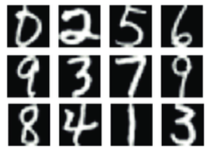
$\hat{Y}^T = n^T \gamma^s Y^S$

Train **classifier** f^T with $(X^T, \hat{Y}^T)$

Predict target labels : $Y^T_{pred} = f^T(X^T)$

Experiments

Settings



USPS ($d = 16 \times 16$, $K = 10$ classes)

$n^S_{train} = 300 \times 10$ or 30×10

nbRep = 10 (random sampling / class)

MNIST ($d = 28 \times 28$, $K = 10$ classes)

$n^T_{train} = 300 \times 10$ or 30×10

nbRep = 10 (random sampling / class)

$n^T_{test} = 200 \times 10$

Number of labelled observations (total: n_* / in each class k: $n_{k,*}$):

dataset	unsupervised	semi-supervised	partial
USPS	$n^S_* = n^S$	$n^S_* = n^S$	$n^S_{k,*} \in \{3;5;25;100\}$
MNIST	$n^T_* = 0$	$n^T_{k,*} \in \{1;3;10\}$	$n^T_{k,*} \in \{3;5;25;100\}$

Classifier f: CNN with 2 convolutional and 2 dense layers

α: 0.01 or 1

Unsupervised and semi-supervised HDA

$n^T_{k,*}$	baseline	$n^S = n^T = 3\ 000$		$n^S = n^T = 300$	
		COOT + LP	JDCOT	COOT+LP	JDCOT
0	-	72.96 ±8.2	77.27 ±9.1	57.27 ±16.2	58.08 ±17.2
1	39.59 ±6.0	75.81 ±4.9	78.45 ±1.1	61.74 ±14.5	69.98 ±2.8
3	56.82 ±4.4	75.35 ±6.5	79.02 ±0.9	69.71 ±7.2	73.19 ±2.4
10	80.49 ±3.1	75.75 ±6.8	88.34 ±1.7	77.25 ±1.7	85.67 ±1.7

Table 1: Mean and standard deviation of the test accuracy (%) over 10 random samplings for the training sets, considering two sample sizes. $n^T_{*,k}$ denotes the number of known labels in each class k, in target domain.
LP = Label Propagation
Baseline = training of f on labelled target data only.

- Improvement w.r.t the baseline score
- Growing performance along with the number of known target labels, even more for smaller sample size
- More stable than CoOT over repetitions

Partial HDA

JDCOT	$n^S_{*,k} = n^T_{*,k}$	3	5	25	100
source	init	70.9 ±4.3	77.9 ±2.3	92 ±0.7	97.6 ±0.4
	final	73.5 ±5.3	84.6 ±2.5	94.6 ±0.9	98 ±0.2
target	init	62.7 ±3.2	70.5 ±3.4	90.2 ±0.9	96.1 ±0.7
	final	68.7 ±5.5	79 ± 3.2	90.3 ±0.5	97 ± 0.2

Table 2: Mean and standard deviation of the test accuracy (%) over 10 random samplings for the training set. $n^S = n^T = 3\ 000$. $n_{*,k}$ denotes the number of known labels in each class k, in each domain. (init) perf. after training on the available target labels, (final) perf. after the whole process

- Improvement of the accuracy both on source and target domains

Conclusion

Conclusion & on-going/future works

- Joint Distribution Co-Optimal Transport (JDCOT) : heterogeneous transfer learning using optimal transport
    - domain adaptation in the case of source and target spaces of different features and different dimensions, matching both samples and features with transport maps and learning the classifier
    - with unsupervised, semi-supervised and partial domain adaptation.
- deep-JDCOT: extension to a deep learning setting (ex. image datasets)
    - simultaneous optimisation of 2 transport plans (samples + variables) and 2 features extractors (source + target)
    - OT between vector representations of the data, optimisation with minibatch stochastic gradient descent
- Different class proportions between source and target data
    - Weakly-supervised strategy
    - Unbalanced / Partial (Co)OT (extra hyper-parameter)

References

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. (2017). Joint distribution optimal transportation for domain adaptation. volume 30.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2016). Optimal Transport for Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865.

Peyré, G. and Cuturi, M. (2018). Computational Optimal Transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–206.

Redko, I., Vayer, T., Flamary, R., and Courty, N. (2020). CO-Optimal Transport. In *Neural Information Processing Systems (NeurIPS)*, Online, Canada.