Unsupervised Domain Adaptation Based on Source-guided Discrepancy

Seiichi Kuroki[1,2] Nontawat Charoenphakdee[1,2] Han Bao[1,2]
Junya Honda[1,2] Issei Sato[1,2] Masashi Sugiyama[2,1]
1: The University of Tokyo 2: RIKEN AIP

Abstract

Proposed a new discrepancy measure for unsupervised domain adaptation.

- Exploits all available information including labels in the source domain unlike existing discrepancy measures.
- Has a tighter generalization error bound and is computationally efficient.

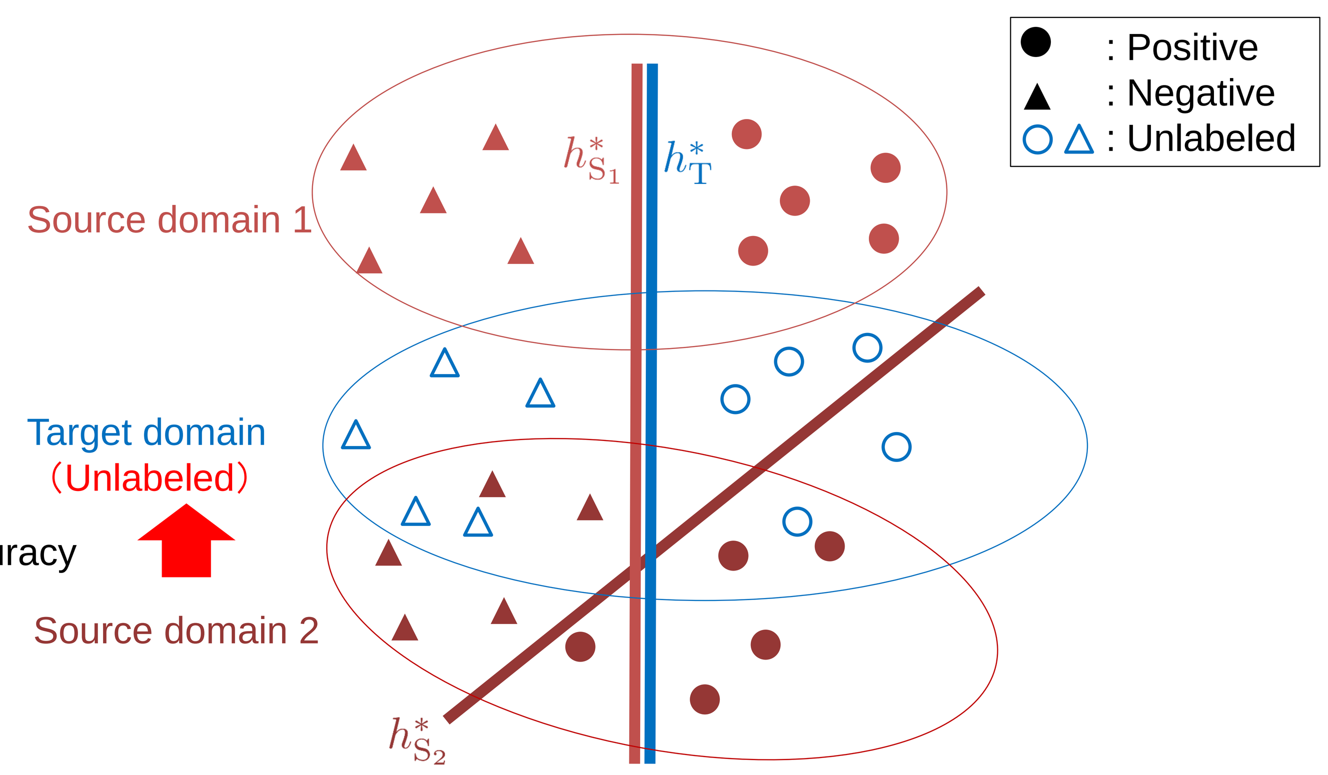
Unsupervised Domain Adaptation

In unsupervised domain adaptation, we consider the following setting.

- Distributions in the source and target domains are related but not the same.
- Labeled data in the source domain and unlabeled data in the target domain.

Ex. Natural language processing, speech recognition, computer vision

How can we utilize labeled data in the source domain?



To select a good source, labels in the source domain might be useful.

Related Work

$\mathcal{X}$-disc [Mansour et al., 2009]

The difference between expected losses of the two domains for the worst pair of hypotheses:

$$\mathrm{disc}^{\ell}_{\mathcal{H}}(P_T, P_S) = \sup_{h,h' \in \mathcal{H}} \left| R^{\ell}_T(h, h') - R^{\ell}_S(h, h') \right|$$

- The computation of $\mathcal{X}$-disc is intractable.

$d_{\mathcal{H}}$ [Ben-David et al., 2010]

A computationally efficient proxy of $\mathcal{X}$-disc:

$$d_{\mathcal{H}}(P_T, P_S) = \sup_{h \in \mathcal{H}} \left| R^{\ell_{01}}_T(h, 1) - R^{\ell_{01}}_S(h, 1) \right|$$

	Using source labels	Generalization error bound	Computation complexity
$\mathcal{X}$-disc	No	Loose	High
$d_{\mathcal{H}}$	No	N/A	Low
S-disc [Proposed]	Yes	Tight	Low

Proposed Measure: Source-guided Discrepancy (S-disc)

Explicitly use the best hypothesis h^*_S in the source domain:

$$\varsigma^{\ell}_{\mathcal{H}}(P_T, P_S) = \sup_{h \in \mathcal{H}} \left| R^{\ell}_T(h, h^*_S) - R^{\ell}_S(h, h^*_S) \right|$$

1. Estimation of h^*_S requires only labeled data in the source domain.
→ Can be estimated from samples.
2. No need to consider a pair of hypotheses.
→ Computationally efficient.
3. The following inequality holds:

$$\left| R^{\ell}_T(h, h^*_S) - R^{\ell}_S(h, h^*_S) \right| \le \varsigma^{\ell}_{\mathcal{H}}(P_T, P_S) \le \mathrm{disc}^{\ell}_{\mathcal{H}}(P_T, P_S)$$

→ Can give a tighter bound than $\mathcal{X}$-disc.

Generalization Error Bound

Theorem 1 If ℓ obeys the triangular inequality,

$$R^{\ell}_T(h, f_T) - R^{\ell}_T(h^*_T, f_T) \le \varsigma^{\ell}_{\mathcal{H}}(P_T, P_S) + R^{\ell}_S(h, h^*_S) + R^{\ell}_T(h^*_S, h^*_T).$$

Regret arising from using h instead of h^*_T

Estimation error in the source domain

Gap of the best classifiers: Uncontrollable in this problem

→ The Lower S-disc, the better generalization.

Estimation Algorithm

S-disc estimation can be reduced to a cost-sensitive classification for the 0-1 loss:

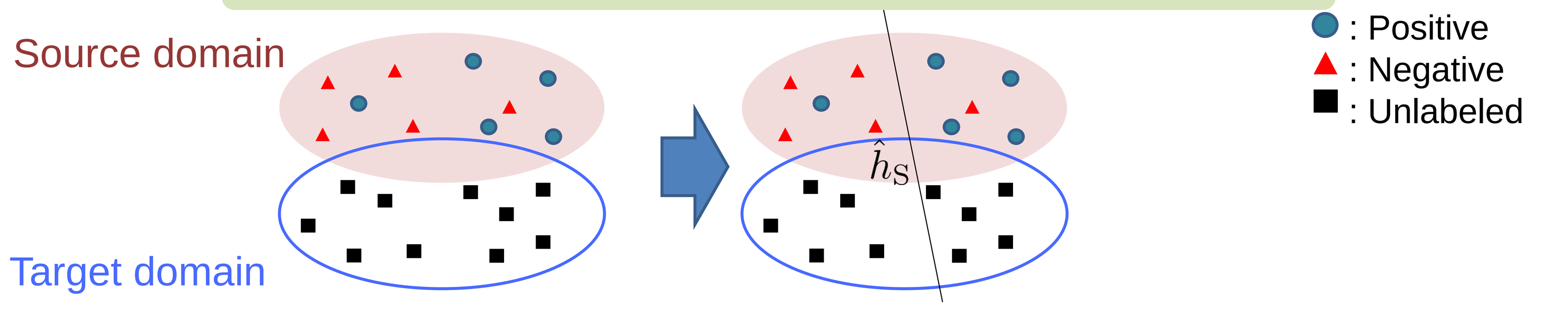
Theorem 2 For a symmetric hypothesis space $\mathcal{H}$ ($h \in \mathcal{H}$ implies $-h \in \mathcal{H}$) ,

$$\varsigma^{\ell_{01}}_{\mathcal{H}}(\widehat{P}_T, \widehat{P}_S) = 1 - \min_{h \in \mathcal{H}} \left[\frac{1}{n_S} \sum_{j=1}^{n_S} \ell(h(x^S_j), h^*_S(x^S_j)) + \frac{1}{n_T} \sum_{i=1}^{n_T} \ell(h(x^T_i), -h^*_S(x^T_i)) \right].$$

Can be expressed as the binary classification loss.

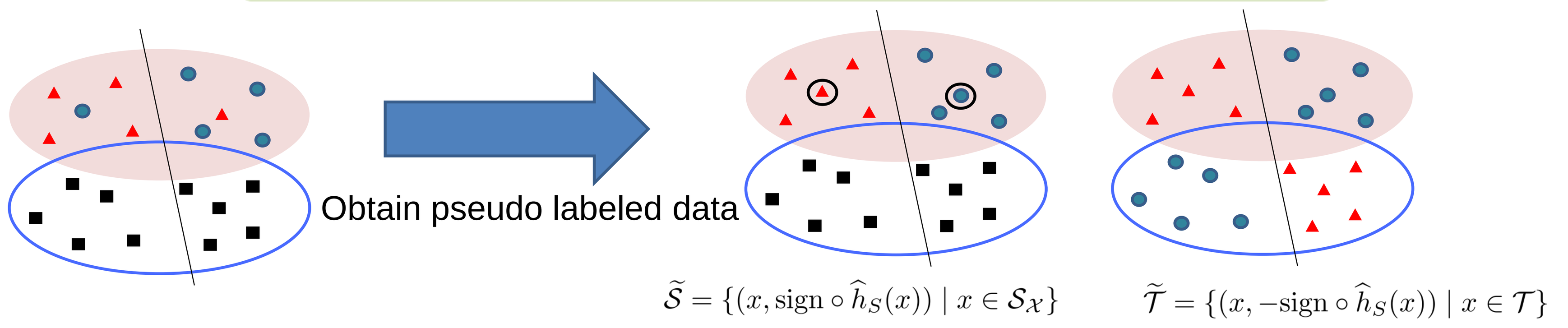
This theorem suggests a three-step algorithm for S-disc estimation described as follows:

Step1: Source Learning

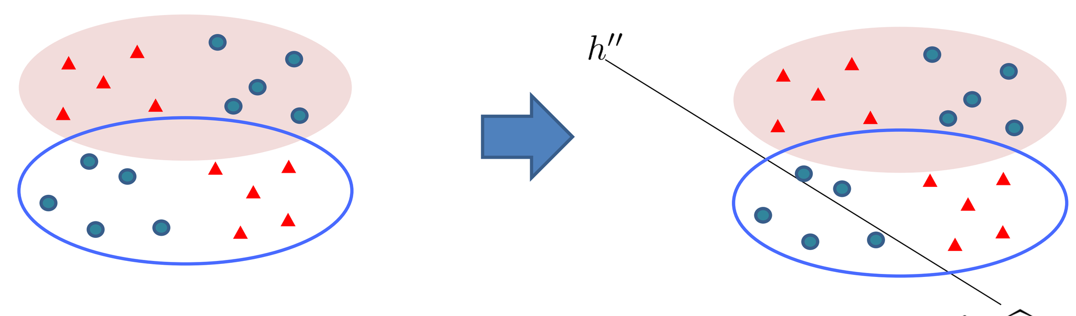


Learn a classifier $\hat{h}_S$ using labeled source data

Step2: Pseudo Labeling



Step3: Cost Sensitive Learning from Pseudo Labeled Data

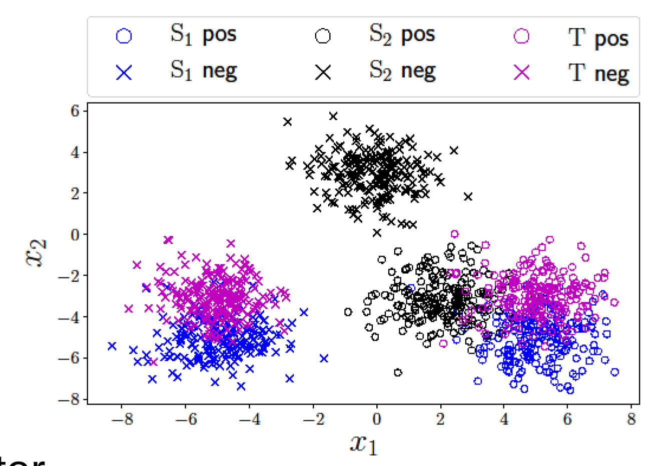


Learn a classifier h'' and obtain $\varsigma^{\ell}_{\mathcal{H}}(\widehat{P}_T, \widehat{P}_S) = 1 - J_{\ell_{01}}(h'')$.

Source Selection Experiments

Toy Dataset

Method: SVM with linear kernel
Data: 200 data points per class for each of two sources S_1, S_2, and target T



We obtain the following results:

$$\varsigma^{\ell}_{\mathcal{H}}(\widehat{P}_T, \widehat{P}_{S_1}) = 0.27, \quad \varsigma^{\ell}_{\mathcal{H}}(\widehat{P}_T, \widehat{P}_{S_2}) = 0.49$$

$$d_{\mathcal{H}}(\widehat{P}_T, \widehat{P}_{S_1}) = 0.69, \quad d_{\mathcal{H}}(\widehat{P}_T, \widehat{P}_{S_2}) = 0.49.$$

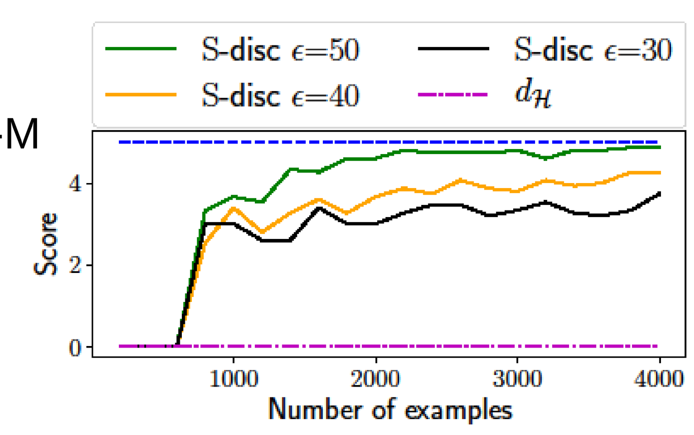
→ S-disc regards S_1 is better while $d_{\mathcal{H}}$ regards S_2 is better.

S-disc is the better discrepancy to measure the quality of sources.

Benchmark Dataset

Method：Logistic Regression
Target：MNIST
Sources: Five Clean MNIST-M and Five Noisy MNIST-M
Task: Classify between even and odd digits

Score = # of clean sources from top 5 sources chosen by each discrepancy measure



1. S-disc achieved a better performance as the number of examples increases.
2. $d_{\mathcal{H}}$ cannot distinguish between noisy and clean sources.

References

[1] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. Machine Learning, 79(1-2):151-175, 2010
[2] Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In COLT, 2009.
[3] Mohri, M., and Medina, A. M. 2012. New analysis and algorithm for learning with drifting distributions. In ALT, 124–138.