Online Transfer Learning with Heterogeneous Source

Yan Yuguang

South China University of Technology

May 12, 2015

Contents

Introduction

Transfer Learning

Transfer learning aims to transfer knowledge extracted from the source domain to the target domain.

- Homogeneous transfer

$$\mathcal{X}^{source} = \mathcal{X}^{target} \; and \; \mathcal{Y}^{source} = \mathcal{Y}^{target}$$

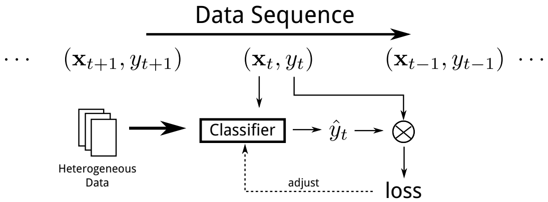
- Heterogeneous transfer

$$\mathcal{X}^{source} \neq \mathcal{X}^{target} \; or \; \mathcal{Y}^{source} \neq \mathcal{Y}^{target}$$

Introduction

Online Heterogeneous Transfer

We investigate online heterogeneous transfer learning problem.



- Heterogeneous transfer

  the feature spaces of the source and target domains are completely different

  for instance, image-text, English-Chinese-French
- Online transfer

  data instances in the target domain arrive sequentially

Introduction

Challenges

- Heterogeneous knowledge transfer across the source and target domains
- No prior training data to build a precise relationship across the source and target domains

Introduction

Online Heterogeneous Transfer Algorithms

Procedure of our proposed OHT algorithms

- construct a connection between the source and target domains via co-occurrence data
- adopt weighted K nearest neighbor algorithm using data in the heterogeneous source
- apply traditional online learning algorithm to train a hypothesis in the target domain
- combine two hypotheses to obtain the ensemble classifier

Related Work

Heterogeneous Transfer

- Build text features for image (Wang, Hoiem, & Forsyth, 2009)
- annotation-based probabilistic latent semantic analysis (Yang, Chen, Xue, Dai, & Yu, 2009)
- Construct representation for image using common semantic view between image and text data (Zhu et al., 2011)
- Co-transfer via transition probability (Ng, Wu, & Ye, 2012; Qingyao Wu, Ng, & Yunming Ye, 2014)

Above-mentioned studies require prior training data in the source and target domains.
In our problem setting, all the instance in the target domain arrive sequentially.

Related Work

Online Learning

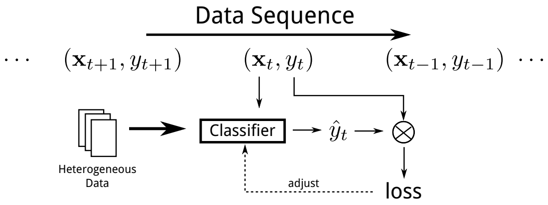
Only a few works consider transfer learning in an online fashion.

- Ensemble learning method for online homogeneous transfer (Zhao & Hoi, 2010; Zhao, Hoi, Wang, & Li, 2014)
- Multi-view method for online heterogeneous transfer

Zhao et al. assume that the feature space of the source domain is a subset of that of the target domain.
In our problem setting, the feature spaces of the source and target domains do not share any common subset (e.g., image and text).

Problem Definition

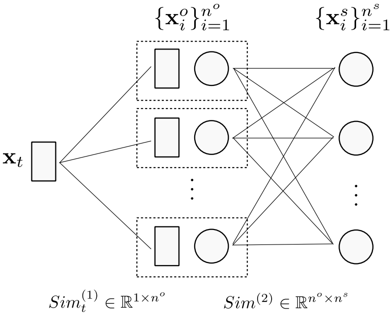
- Given some instances $\{\mathbf{x}_i^s, y_i^s\}_{i=1}^{n^s}$ in the source data space $\mathcal{X}^s \times \mathcal{Y}^s$, where $\mathcal{X}^s = \mathbb{R}^{d^s}$ and $\mathcal{Y}^s = \{+1, -1\}$.
- The data space of the target domain is $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{+1, -1\}$.
- $\mathcal{Y} = \mathcal{Y}^s$, and $\mathcal{X} \cap \mathcal{X}^s = \varnothing$.



The objective of online heterogeneous transfer is to learn a prediction function $f(\mathbf{x}_t)$ to classify the instance on the target domain in an online fashion.

Methods

Heterogeneous Knowledge Transfer



Similarity between $\mathbf{x}_t$ and heterogeneous instances

$$Sim_t(j) = \sum_i Sim_t^{(1)}(i) Sim^{(2)}(i,j)$$

Hypothesis

$$h_t^s(\mathbf{x}_t) = \sum_{i\in N} \frac{Sim_t(i)}{\sum_{i\in N} Sim_t(i)} y_i$$

where N is the set of identifiers of K nearest neighbors in the source

Online Heterogeneous Transfer Algorithms

OHT1

Algorithm 1 OHT

```
Input:  aggressiveness parameter $C > 0$
        $\eta = \frac{1}{2}$
        heterogeneous source data
Output: $\mathbf{v} = \mathbf{0}$, $w_1^s \in (0,1)$, $w_1 \in (0,1)$, where $w_1^s + w_1 = 1$
for $t = 1$ to $T$ do
    receive instance: $\mathbf{x}_t \in \mathcal{X}$
    normalize: $\theta_t^s = \frac{w_t^s}{w_t^s + w_t}, \theta_t = \frac{w_t}{w_t^s + w_t}$
    predict: $\hat{y}_t = \text{sign}\left(\theta_t^s \Omega(h_t^s(\mathbf{x}_t)) + \theta_t \Omega(\mathbf{v}_t \cdot \mathbf{x}_t) - \frac{1}{2}\right)$
    receive correct label: $y_t \in \mathcal{Y}$
    compute:
        $$w_{t+1}^s = w_t^s \exp\left\{-\eta\left(\Omega(h_t^s(\mathbf{x}_t)) - \Omega(y_t)\right)^2\right\}$$
        $$w_{t+1} = w_t \exp\left\{-\eta\left(\Omega(\mathbf{v}_t \cdot \mathbf{x}_t) - \Omega(y_t)\right)^2\right\}$$
    suffer loss: $\ell_t^* = \max\{0, 1 - y_t(\mathbf{v}_t \cdot \mathbf{x}_t)\}$
    set: $\tau_t = \frac{\ell_t^*}{\|\mathbf{x}_t\|^2}$   (I: $\tau_t = \min\{C, \frac{\ell_t^*}{\|\mathbf{x}_t\|^2}\}$, II: $\tau_t = \frac{\ell_t^*}{\|\mathbf{x}_t\|^2 + \frac{1}{2C}}$ )
    update: $\mathbf{v}_{t+1} = \mathbf{v}_t + \tau_t y_t \mathbf{x}_t$
end for
```

Project function $\Omega(z) = \max\{0, \min\{1, \frac{z+1}{2}\}\}$.

Online Heterogeneous Transfer Algorithms

OHT2

Algorithm 2 OHT2

```
Input:  aggressiveness parameter $C > 0$
        discount parameter $\alpha \in (0, 1)$
        heterogeneous source data
Output: $\mathbf{v} = \mathbf{0}$, $w_1^s \in (0, 1)$, $w_1 \in (0, 1)$, where $w_1^s + w_1 = 1$
for $t = 1$ to $T$ do
    receive instance: $\mathbf{x}_t \in \mathcal{X}$
    normalize: $\theta_t^s = \frac{w_t^s}{w_t^s + w_t}, \theta_t = \frac{w_t}{w_t^s + w_t}$
    predict: $\hat{y}_t = \mathrm{sign}\left(\theta_t^s \, \mathrm{sign}(h_t^s(\mathbf{x}_t)) + \theta_t \, \mathrm{sign}(\mathbf{v}_t \cdot \mathbf{x}_t)\right)$
    receive correct label: $y_t \in \mathcal{Y}$
    compute:
        $w_{t+1}^s = w_{t+1}^s \alpha^{I(y_t h_t^s(\mathbf{x}_t) \leq 0)}$
        $w_{t+1} = w_{t+1} \alpha^{I(y_t(\mathbf{v}_t \cdot \mathbf{x}_t) \leq 0)}$
    suffer loss: $\ell_t^* = \max\{0, 1 - y_t(\mathbf{v}_t \cdot \mathbf{x}_t)\}$
    set: $\tau_t = \frac{\ell_t^*}{\|\mathbf{x}_t\|^2}$   (I: $\tau_t = \min\{C, \frac{\ell_t^*}{\|\mathbf{x}_t\|^2}\}$, II: $\tau_t = \frac{\ell_t^*}{\|\mathbf{x}_t\|^2 + \frac{1}{2C}}$ )
    update: $\mathbf{v}_{t+1} = \mathbf{v}_t + \tau_t y_t \mathbf{x}_t$
end for
```

Indication function

$$I(z) = \begin{cases} 1, & z = TRUE \\ 0, & z = FALSE \end{cases}$$

Theoretical Analysis

$\mathbf{Hedge}(\beta)$ Algorithm

At t-th trial, $\mathbf{Hedge}(\beta)$ algorithm synthesizes opinions from different experts based on a weight vector $\mathbf{weight}_t$, and updates the weight vector using rule

$$weight_{t+1}^{i} = weight_{t}^{i} \cdot \beta^{loss_t^i}$$

where $\beta \in [0, 1]$ and $loss_t^i \in [0, 1]$.

$\mathbf{Hedge}(\beta)$	OHT1	OHT2
$loss_t^i$	$\ell_t^s = (\Omega(h_t^s(\mathbf{x}_t)) - \Omega(y_t))^2$ $\ell_t = (\Omega(\mathbf{v}_t \cdot \mathbf{x}_t) - \Omega(y_t))^2$	$\ell_t^s = I(y_t h_t^s(\mathbf{x}_t) \leq 0)$ $\ell_t = I(y_t(\mathbf{v}_t \cdot \mathbf{x}_t) \leq 0)$
β	$\beta = \exp\{-\eta\}$	$\beta = \alpha$

OHT algorithms obey the rule of $\mathbf{Hedge}(\beta)$ algorithm.

Theoretical Analysis

Proposition

Proposition 1

Given loss $\ell_t^s \in [0,1]$, $\ell_t \in [0,1]$ *and decay factor* $\beta \in (0,1)$, *for any sequence of loss vectors* $\{(\ell_t^s, \ell_t) | t = 1, 2, \cdots T\}$, *we have*

$$\sum_{t=1}^{T} (\theta_t^s \ell_t^s + \theta_t \ell_t) \leq \frac{1}{1-\beta} \min\left(\Delta^s, \Delta\right)$$

where $\Delta^s = \ln(\frac{1}{w_1^s}) + (\ln\frac{1}{\beta}) \sum_{t=1}^{T} \ell_t^s$ *and* $\Delta = \ln(\frac{1}{w_1}) + (\ln\frac{1}{\beta}) \sum_{t=1}^{T} \ell_t$

Theoretical Analysis

Theorem of OHT1

Theorem 1 (Mistake bound of OHT1)

Let M be the number of mistakes made by OHT1 algorithm, then we have

$$M \leq \frac{4}{1-\beta} \min(\Delta^s, \Delta)$$

where $\Delta^s = \ln(\frac{1}{w_1^s}) + (\ln \frac{1}{\beta}) \sum_{t=1}^{T} \ell_t^s$ *and* $\Delta = \ln(\frac{1}{w_1}) + (\ln \frac{1}{\beta}) \sum_{t=1}^{T} \ell_t$.

Theoretical Analysis
Theorem of OHT2

Theorem 2 (Mistake bound of OHT2)

Let M be the number of mistakes made by OHT2 algorithm, then we have

$$M \leq \frac{2}{1-\beta} \min(\Delta^s, \Delta)$$

where $\Delta^s = \ln(\frac{1}{w_1^s}) + (\ln \frac{1}{\beta}) \sum_{t=1}^{T} \ell_t^s$ *and* $\Delta = \ln(\frac{1}{w_1}) + (\ln \frac{1}{\beta}) \sum_{t=1}^{T} \ell_t$.

Theorem 3 (Recommended value of β)

When

$$\beta = \frac{\sqrt{T}}{\sqrt{T} + \sqrt{2 \max(\ln(\frac{1}{w_1^s}), \ln(\frac{1}{w_1}))}}$$

we have

$$M \leq 2 \min(\Lambda^s, \Lambda) + \sqrt{2T \max(\ln(\frac{1}{w_1^s}), \ln(\frac{1}{w_1}))}$$

where $\Lambda^s = \sum_{t=1}^{T} \ell_t^s + \ln(\frac{1}{w_1^s})$ *and* $\Lambda = \sum_{t=1}^{T} \ell_t + \ln(\frac{1}{w_1})$.

Experiments

Dataset

NUS-WIDE dataset

- Target domain: Image
- Heterogeneous source: Text
- Co-occurrence data: co-occurred image-tag pairs

Experiments

Baseline Methods

- Passive-Aggressive algorithms

  Traditional online learning algorithm
- Kernel function

  Gaussian Kernel
- Number of nearest neighbors

  $K = 100$

Experiments

Results

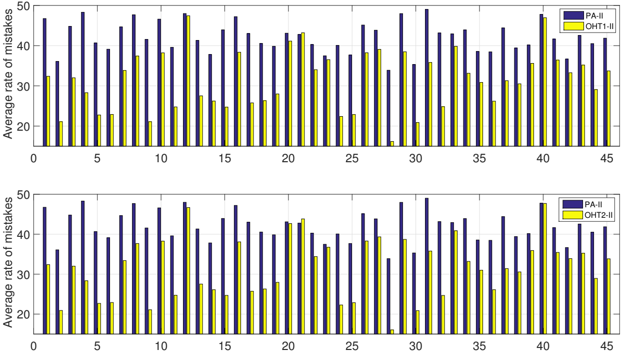


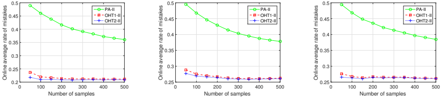
Figure : Average rate of mistakes on all 45 tasks

Observations:

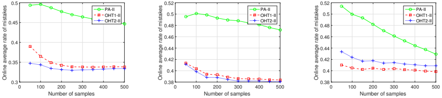
- The mistake rate of PA-II is very high.
- OHT algorithms generally outperform PA-II.

Experiments

Results



(a) Task 2 (b) Task 14 (c) Task 36



(d) Task 7 (e) Task 16 (f) Task 33

Figure : Online average rate of mistakes on example tasks

Observations:

- OHT algorithms usually achieve better performance at the beginning stage.
- On some tasks (e.g., 7, 16 and 33), the mistake rates of all algorithms decrease, but OHT methods always perform better.

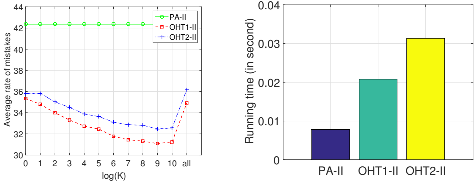
Experiments

Significant Test

- Paired t-test ($\alpha = 0.01$)
  OHT1 vs. PA: 44/0/1
  OHT2 vs. PA: 42/2/1
- Cohen's d value
  ($d > 0.8$: large promotion, $d \in (0.2, 0.8)$: middle promotion)
  OHT1: 41/3
  OHT2: 40/3

Experiments

Parameters and Running Time



(a) (b)

Figure : (a) The average rate of mistakes under varying values of K. (b) The average running time of different algorithms when all instances in heterogeneous source are considered.

Observations:

- OHT algorithms consistently outperform PA
- By using global nearest neighbor approach, we can obtain generally comparable running time to PA, and achieve a better performance than PA.

Conclusion and Future Works

Conclusion

- We explore online heterogeneous transfer learning problem.
- We construct a connection across the domains using co-occurrence data, and apply the ensemble strategy to train a classifier.
- We offer the theoretical analysis of our algorithms.
- Experimental results show the effectiveness of our algorithms.

Future works:

- applications with other types of data
- multiple source domains

References I

Ng, M. K., Wu, Q., & Ye, Y. (2012). Co-transfer learning via joint transition probability graph based method. In *Proceedings of the 1st international workshop on cross domain knowledge discovery in web and social network mining* (pp. 1–9).

Qingyao Wu, Ng, M., & Yunming Ye. (2014, July). Cotransfer Learning Using Coupled Markov Chains with Restart [Journal Paper]. *IEEE Intelligent Systems*, *29*, 26-33.

Wang, G., Hoiem, D., & Forsyth, D. (2009). Building text features for object image classification. In *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on* (pp. 1367–1374).

References II

Yang, Q., Chen, Y., Xue, G.-R., Dai, W., & Yu, Y. (2009). Heterogeneous transfer learning for image clustering via the social web. In *Proceedings of the joint conference of the 47th annual meeting of the acl and the 4th international joint conference on natural language processing of the afnlp: Volume 1-volume 1* (pp. 1–9).

Zhao, P., & Hoi, S. C. (2010). Otl: A framework of online transfer learning. In *Proceedings of the 27th international conference on machine learning (icml-10)* (pp. 1231–1238).

Zhao, P., Hoi, S. C., Wang, J., & Li, B. (2014). Online transfer learning. *Artificial Intelligence*, *216*, 76–102.

Zhu, Y., Chen, Y., Lu, Z., Pan, S. J., Xue, G.-R., Yu, Y., & Yang, Q. (2011). Heterogeneous transfer learning for image classification. In *Aaai*.

THANK YOU
FOR
YOUR ATTENTION!