Information-Theoretical Learning of Discriminative Clusters for Unsupervised Domain Adaptation

Yuan Shi   Fei Sha

University of Southern California

Appears in ICML 2012
Presented by Liming Wang
Introduction

- Domain adaption (aka. transfer learning): How to deal with mismatch between training data and the test data.
  - Example: Face detection system for Facebook mobile users using data captured on webcams.
- Unsupervised domain adaptation: no labels on the target data.
  - Challenging: Target does not explicitly provide any information on how to optimize classifiers.
Previous works

  - Shift the labeled data to the target domain and train the classifiers.

- Augment original features so that the new features are domain invariant (Pan et al, IEEE TNN, 2011; Gopalan et al, ICCV, 2011; Glorot et al, ICML 2011).

- Underlying assumptions of all the methods: there exists a domain-invariant feature space such that the marginal distributions are the same.
  - Same marginals do not necessarily imply same posteriors. One need to find the discriminative one.
Problem statement

- $N$ labeled instances from the source domain $\{(x_s, y_s)\}$ where $x_s \in \mathcal{X} \subset \mathbb{R}^D$ and $y_s \in \mathcal{Y} = \{1, \ldots, C\}$.
- $M$ unlabeled instance from the target domain $\{x_t\}$ where $x_t \in \mathcal{X}$.
- Goal: Find a classifier $f : x \in \mathcal{X} \mapsto y \in \mathcal{Y}$ which performs well on the target domain.
- The problem is ill-posed. We will use the following key assumptions.
Discriminative clustering

Key Assumptions: There exists a feature space such that we have

1. **Separation**: Data in the source domain and target domain are tightly clustered.

2. **Alignment**: The clusters from two domains correspond to the same label are geometrically close.

![Diagram showing the concept of separation and alignment in discriminative clustering.](image-url)
Conditional models in the feature space

- Consider the feature space of dimension $d$ and use the linear transformation $L \in \mathbb{R}^{d \times D}$. Moreover in the new feature space, we use $1 - \text{NN}$ to classify (Recall key assumption 1).
- Define a (Mahalanobis) distance between $x_i$ and $x_j$ as

$$d_{ij}^2 := \| Lx_i - Lx_j \|^2_2.$$

- Define the conditional probability of having $x_j$ as the nearest neighbor of $x_i$ as

$$p_{ij} := \frac{e^{-d_{ij}^2}}{\sum_{j \neq i} e^{-d_{ij}^2}}.$$

- The posterior can be estimated as

$$\hat{p}_{ic} := p(y_i = c|x_i) = \sum_{j \neq i} p_{ij} \delta_{jc},$$

if we know all the labels of $x_j$. 
Discriminative clustering in the source and target

- If we know the label for $x_i$, then $\sum_c \hat{p}_{ic}\delta_{ic}$ is the probability of correctly classifying $x_i$.
- We can use the data from source as the regularizer, i.e. minimizing the classifying error:

$$\epsilon_s := 1 - \frac{1}{N} \sum_s \sum_c \hat{p}_{sc}\delta_{sc}$$

- Because of key assumption 2, $\hat{p}_{tc}$ using all available source labels should be closed to the true posterior.
- Therefore, the posterior $\hat{p}_t := [\hat{p}_{t1}, \ldots, \hat{p}_{tC}]$ should resemble a delta function.
We will maximize the mutual information between the data and the estimated label $\hat{Y}$

$$l_t(X; \hat{Y}) = H[p_0] - \frac{1}{M} \sum_t H[p_t].$$

- $p_0 := \frac{1}{M} \sum_t p_t$

- Minimizing directly $\sum_t H(p_t)$ may lead to degenerated case.
Discriminability between source and target

- The source and target should share some common probability supports in the feature space.
- Consider a binary classifier for $x_i$. Define $q_i := \delta_{x_i \in S}$. Similar to previous case, we have the estimated posterior as before

$$\hat{p}_{ic} := p(y_i = c | x_i) = \sum_{j \neq i} p_{ij} \delta_{jc},$$

- We need to minimize the mutual information

$$I_{st}(X; Q) = H[\hat{q}_0] - \frac{1}{M+N} \sum_i H[\hat{q}_i].$$

- idea: The feature space should be discriminative between source and target such that one has the least information from other instances.
Learning and model selection

- The problem can be formulated as the following optimization problem

\[ L^* = \arg \min ( - I_t(X; \hat{Y}) + \lambda I_{st}(X; Q)) \]
\[ \text{s.t.} \quad \text{tr}(L^T L) \leq d \]

- \( \lambda \) is selected by cross-validation for \( \epsilon_s \).

- The optimization is solved by gradient descent.
Experiments

- Four databases: Caltech-256, Amazon, Webcam (low-res.), DSLR (high-res.).

<table>
<thead>
<tr>
<th>Domains</th>
<th>PCA</th>
<th>TCA</th>
<th>GFS</th>
<th>LMNN</th>
<th>Metric</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSLR $\rightarrow$ Webcam</td>
<td>80.6±0.5</td>
<td>66.2±0.5</td>
<td>75.5±0.4</td>
<td>81.3±0.4</td>
<td>55.6±0.7</td>
<td>83.6±0.5</td>
</tr>
<tr>
<td>DSLR $\rightarrow$ Amazon</td>
<td>35.1±0.3</td>
<td>31.4±0.2</td>
<td>35.7±0.5</td>
<td>42.3±0.3</td>
<td>30.3±0.8</td>
<td>39.6±0.4</td>
</tr>
<tr>
<td>Caltech $\rightarrow$ DSLR</td>
<td>36.6±1.2</td>
<td>33.1±0.8</td>
<td>36.5±0.9</td>
<td>37.2±1.1</td>
<td>35.0±1.1</td>
<td>44.4±1.2</td>
</tr>
<tr>
<td>Caltech $\rightarrow$ Amazon</td>
<td>37.7±0.5</td>
<td>34.9±0.4</td>
<td>37.9±0.5</td>
<td>43.2±0.4</td>
<td>33.7±0.8</td>
<td>49.2±0.6</td>
</tr>
<tr>
<td>Amazon $\rightarrow$ Webcam</td>
<td>33.1±0.6</td>
<td>26.5±0.8</td>
<td>32.8±0.7</td>
<td>35.2±0.8</td>
<td>36.0±1.0</td>
<td>38.5±1.3</td>
</tr>
<tr>
<td>Amazon $\rightarrow$ Caltech</td>
<td>35.9±0.3</td>
<td>29.3±0.3</td>
<td>36.1±0.5</td>
<td>37.6±0.4</td>
<td>27.3±0.7</td>
<td>40.0±0.4</td>
</tr>
</tbody>
</table>