Domain Generalization via Invariant Feature Representation

Krikamol Muandet\textsuperscript{1}, David Balduzzi\textsuperscript{2}, Bernhard Schölkopf\textsuperscript{1}

\textsuperscript{1}Empirical Inference Department, MPI for Intelligent Systems
\textsuperscript{2}Machine Learning Laboratory, ETH Zurich

June 18, 2013
Flow Cytometry

Image courtesy of American Journal of Clinical Pathology.

Domains = Patients ($\mathbb{P}_{XY}$), Train Data $\{X_j^{(i)}, Y_j^{(i)}\}_{j=1}^{n}$. 
Domain Generalization

\[ P \]

\[ P_{XY}^1 \quad P_{XY}^2 \quad \ldots \quad P_{XY}^N \quad P_X \]

\[ (X_k, Y_k) \quad (X_k, Y_k) \quad \ldots \quad (X_k, Y_k) \quad X_k \]

\[ k = 1, \ldots, n_1 \quad k = 1, \ldots, n_2 \quad k = 1, \ldots, n_N \quad k = 1, \ldots, n \]

training data

unseen test data
Related Works

Domain Adaptation (Bickel, Brückner, and Scheffer 2009)

Deal with a mismatch between training and test distributions.
Related Works

Multitask Learning (Caruana 1997)

Learn multiple tasks simultaneously.
Domain Generalization
Blanchard, Lee, and Scott 2011

Generalize from multiple source domains to previously unseen domains.
Domain Generalization

Problem Setting

Train: The joint distributions $P^1_{XY}, P^2_{XY}, \ldots, P^N_{XY} \sim P$.

Prediction: An unseen distribution $P^*_{X} \sim P$.

Goal: Learn $f : \mathcal{X} \times X^* \rightarrow Y$.

Assume: $P^1_{Y|X} \approx P^2_{Y|X} \approx \cdots \approx P^N_{Y|X}$.

i.e. functional relationship is stable
Domain Generalization

Problem Setting

Train: The joint distributions $P^1_{XY}, P^2_{XY}, \ldots, P^N_{XY} \sim \mathcal{P}$.

Prediction: An unseen distribution $P^*_X \sim \mathcal{P}$.

Goal: Learn $f : \mathcal{X} \times \mathcal{X}^* \rightarrow \mathcal{Y}$.

Assume: $P^1_{Y|X} \approx P^2_{Y|X} \approx \cdots \approx P^N_{Y|X}$.

i.e. functional relationship is stable

Domain Adaptation under Target and Conditional Shift
K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang (ICML2013)
Objective

Find feature representation, $\mathcal{B}(X)$ that is *invariant* across domains.

1. minimize the distance between empirical distributions $\hat{P}_1^X, \hat{P}_2^X, \ldots, \hat{P}_N^X$ of the transformed samples $\mathcal{B}(X)$.

2. preserve functional relationship between $X$ and $Y$. 

\[
\begin{align*}
\mathbb{P}_1^{Y|X} \cdot \mathbb{P}_1^X & \quad \mathbb{P}_2^{Y|X} \cdot \mathbb{P}_2^X & \quad \cdots & \quad \mathbb{P}_N^{Y|X} \cdot \mathbb{P}_N^X
\end{align*}
\]
Minimizing Distributional Variance

Hilbert space embedding

\[ \mu : \mathcal{P}_X \rightarrow \mathcal{H}, \quad \mathbb{P} \mapsto \int_X k(x, \cdot) \, d\mathbb{P}(x) =: \mu_{\mathbb{P}}. \]
Minimizing Distributional Variance

Find transformation $B$ that minimizes

$$\nabla_H(P) = \frac{1}{N} \sum_{i=1}^{N} ||\mu_i B - \bar{\mu} B||_H^2$$
Minimizing Distributional Variance

- Minimizing distributional variance **alone** does not necessarily help with generalization!
  - Setting $B = 0$ gives zero distributional variance!

- We **also** need to preserve the functional relationship between $X$ and $Y$ encoded in $\mathbb{P}_{Y|X}$. 
Preserving Functional Relationship

Central Subspace
The central subspace $C$ is the minimal subspace that captures the functional relationship between $X$ and $Y$, i.e. $Y \perp \perp X | C^T X$.

Theorem (Li 1991; Kim and Pavlovic 2011; Muandet 2013)
If $B$ maximizes

$$b_k^T \Sigma_{xx}^{-1} \nabla \left( \mathbb{E}[X | Y] \right) \Sigma_{xx} b_k$$

then $Y \perp \perp X | B^T X$. 


Domain-Invariant Component Analysis

\[ \max_{B \in \mathbb{R}^{n \times m}} \frac{f(B)}{g(B)} \]

preserve functional relationship.

minimize distributional variance.
Domain-Invariant Component Analysis

preserve functional relationship.

\[ b_k^T \hat{\Sigma}_{xx}^{-1} \hat{\nabla}(\mathbb{E}[X|Y]) \hat{\Sigma}_{xx} b_k \]

max

\[ f(B) \]

g(B)

minimize distributional variance.

\[ \nabla_{\mathcal{H}}(P) = \sum_{i=1}^{N} \| \mu_i B - \bar{\mu}_B \|_{\mathcal{H}}^2 \]
Domain-Invariant Component Analysis

preserve functional relationship.

\[ b_k^\top \hat{\Sigma}_{xx}^{-1} \hat{V}(\mathbb{E}[X|Y]) \hat{\Sigma}_{xx} b_k \]

\[
\max_{B \in \mathbb{R}^{n \times m}} \frac{1}{n} \text{tr} \left( B^\top L (L + n\varepsilon I_n)^{-1} K^2 B \right)
\]

\[ \text{tr} (B^\top KQKB + BKB) \]

minimize distributional variance.

\[ \nabla \mathcal{H}(\mathcal{P}) = \sum_{i=1}^{N} \| \mu_i B - \bar{\mu} B \|_H^2 \]
Domain-Invariant Component Analysis

Maximization Problem

\[
\max_{B \in \mathbb{R}^{n \times m}} \frac{1}{n} \text{tr} \left( B^T (L + n\epsilon I_n)^{-1} K^2 B \right)
\]

\[
\text{tr} \left( B^T KQKB + BKB \right)
\]

\[
\Downarrow
\]

Generalized Eigenvalue Problem

\[
\frac{1}{n} L(L + n\epsilon I)^{-1} K^2 B = (KQK + K + \lambda I)B \Gamma
\]
Learning guarantee

**Theorem**

Under reasonable assumptions, it holds with probability at least $1 - \delta$ that,

\[
\mathbb{E}[\text{error}] \leq c_1 \mathbb{V}_H(\mathcal{P} \cdot \mathcal{B}) + L(n, N).
\]

- Bound depends on the distributional variance.
- $L(n, N) \to 0$ as samples $n$ and domains $N$ go to infinity.
Experimental Results

**Synthetic Data**

- Generate 10 collections of $n_i \sim \text{Poisson}(200)$ data points.
- For each collection, $x \sim \mathcal{N}(0, \Sigma_i)$ where $\Sigma_i \sim \mathcal{W}(0.2 \times I_5, 10)$.
- The output value is $y = \text{sign}(b_1^T x + \varepsilon_1) \cdot \log(|b_2^T x + c + \varepsilon_2|)$, where $\varepsilon_1, \varepsilon_2 \sim \mathcal{N}(0, 1)$.
Experimental Results: synthetic data

COIR

DICA
Experimental Results

Real-world Data
- Flow cytometry dataset (classification).
- Parkinson’s telemonitoring dataset (regression).

Learning algorithms
- **Pooling SVM**: pool data from all domains and apply standard SVM.
- **Distributional SVM**: apply the kernel

\[
\kappa((\mathbb{P}^i, x^i_k), (\mathbb{P}^j, x^j_l)) = K(\mathbb{P}^i, \mathbb{P}^j) \cdot k(x^k_i, x^l_j)
\]

(Blanchard, Lee, and Scott 2011).
### Experimental Results: Flow cytometry

<table>
<thead>
<tr>
<th>Methods</th>
<th>Pooling SVM</th>
<th>Distributional SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>92.03±8.21</td>
<td>93.19±7.20</td>
</tr>
<tr>
<td>KPCA</td>
<td>91.99±9.02</td>
<td>93.11±6.83</td>
</tr>
<tr>
<td>COIR</td>
<td>92.40±8.63</td>
<td>92.92±8.20</td>
</tr>
<tr>
<td>UDICA</td>
<td>92.51±5.09</td>
<td>92.74±5.01</td>
</tr>
<tr>
<td>DICA</td>
<td><strong>92.72±6.41</strong></td>
<td><strong>94.80±3.81</strong></td>
</tr>
</tbody>
</table>

Similar results for Parkinson’s telemonitoring dataset.
Domain-Invariance Component Analysis (DICA) finds an invariant representation that

- minimizes “differences” between domains
- while preserving discriminative information.

To learn more, please come to our poster!
Thank you!