Flexible Domain Adaptation for Multimedia Indexing

Emilie Morvant, Amaury Habrard, Stéphane Ayache
(emilie.morvant,amaury.habrard,stephane.ayache)@lip.univ-mrs.fr

INTRODUCTION, NOTATIONS AND MOTIVATION

We consider binary classification task:
- $X$ input space, $Y = \{-1, 1\}$ label set
- $P_s$, source domain: distribution over $X \times Y$
- $P_t$, target domain: distribution over $X \times Y$
- $D_s$, marginal distribution over $X$
- $D_t$, marginal distribution over $X$
- $\phi(h)$: target domain: different distribution over $X \times Y$
- $\epsilon_{err}(h)$, target domain errors
- $\epsilon_{err}(h)$, target domain errors
- Supervised Classification objective: $h \in H$ with a low $\epsilon_{err}(h)$
- Domain Adaptation objective: $h \in H$ with a low $\epsilon_{err}(h)$

For example:
- We have labeled images from a Web image corpus, i.e. $P_s$
- Is there a Person in unlabeled images from a Video corpus, i.e. $D_t$?

$\Rightarrow$ We have labeled images from a Web image corpus, i.e. $P_s$
$\Rightarrow$ Is there a Person in unlabeled images from a Video corpus, i.e. $D_t$?

$\Rightarrow$ The Learning distribution is different from the Testing distribution
$\Rightarrow$ How can we learn, from the source domain, a low-error classifier on the target domain?

DOMAINE ADAPTATION

Theorem 1 ([2]). Let $H$ an hypothesis space. If $D_s$ and $D_t$ are respectively the marginal distributions of source and target instance, then for all $\delta \in [0, 1]$, with probability at least $1-\delta$, for every $h \in H$:

$$\epsilon_{\text{err}}(h) \leq \epsilon_{\text{err}}(h) + \frac{1}{2}d_{\text{Hadamard}}(D_s, D_t) + \nu,$$

where $d_{\text{Hadamard}}(D_s, D_t)$ is the Hadamard distance between $D_s$ and $D_t$ and $\nu = \epsilon_{\text{err}}(h') + \epsilon_{\text{err}}(h)$, with $h' = \text{argmin}_{h \in H} \epsilon_{\text{err}}(h) + \epsilon_{\text{err}}(h)$.

**Definition 1 ([1]), $K : X \times X \rightarrow [-1; 1]$ is an $(\epsilon, \tau, \gamma)$-good similarity function for a binary classification problem $P$ if

(i) $A 1 - \epsilon$ probability mass of examples $(x, y)$ satisfy

$$\mathbb{E}_{(x', y') \sim P_s} [\text{sgn}(K(x, x') - R(x'))] \geq \gamma.$$  

(ii) $\mathbb{P}_{x \sim D_t} [R(x')] \geq \tau$ (Notation: $R$ set of reasonable points).

**Properties**
- Generalization of kernels: $K$ may be not symmetric and not PSD
- A low-error linear classifier can be learned by minimizing the Pb. $(S \phi)$ in the explicit projection space defined by $R = \{x_1, \ldots, x_k\}$,

$$\phi^R(.) = \{K(x, x_1), \ldots, K(x, x_k)\}$$

(Notation: $H_{\text{SF}}$ the hypothesis space of such classifiers)

Learning with Good Similarity Functions

**Definition 1 ([1]), $K : X \times X \rightarrow [-1; 1]$ is an $(\epsilon, \tau, \gamma)$-good similarity function for a binary classification problem $P$ if

(i) $A 1 - \epsilon$ probability mass of examples $(x, y)$ satisfy

$$\mathbb{E}_{(x', y') \sim P_s} [\text{sgn}(K(x, x') - R(x'))] \geq \gamma.$$  

(ii) $\mathbb{P}_{x \sim D_t} [R(x')] \geq \tau$ (Notation: $R$ set of reasonable points).

**Properties**
- Generalization of kernels: $K$ may be not symmetric and not PSD
- A low-error linear classifier can be learned by minimizing the Pb. $(S \phi)$ in the explicit projection space defined by $R = \{x_1, \ldots, x_k\}$,

$$\phi^R(.) = \{K(x, x_1), \ldots, K(x, x_k)\}$$

(Notation: $H_{\text{SF}}$ the hypothesis space of such classifiers)

**Some little results**
- Sparsity Analysis. With $\beta_B = \min_{x \sim P_s} \max_{x' \sim P_t} |K(x, x')| K(x, x')^2$, $M_\eta$ is the $\eta$-covering number of $\mathcal{H}$.

$$\left| \langle \alpha', \phi \rangle \right| \leq \frac{1}{\beta_B + \lambda}$$

- Generalization bound. Following the robustness notion of XuKraMann [3], the problem ($\text{DASF}$) is $(2M_\eta, \eta)$ robust on $P_s$, with $\eta > 0$, $M_B_\eta$ is the $\eta$-covering number of $\mathcal{H}$.

$$\epsilon_{\text{err}}(h) \leq \epsilon_{\text{err}}(h) + \frac{1}{\beta_B + \lambda} \sum_{\mathbb{P}_{x \sim D_t} R(x') \geq \tau} \frac{M_\eta}{d_i} + \frac{1}{2}d_{\text{Hadamard}}(D_s, D_t) + \nu.$$