Analysis of Representations for Domain Adaptation

by Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira
presented by Marina Sokolova

Best-of-NIPS'2006

Introduction

- **Domain** is a distribution $\mathcal{D}$ on an instance set $\mathcal{X}$
- Domain adaptation of a classifier
  - A classification task
  - *Source* domain ($\mathcal{D}_S$)
  - *Target* domain ($\mathcal{D}_T$)
  - **Different** distributions
  - Labelled data in *source* domain
  - Unlabelled data in *both* domains
- Examples: spam filters, parsing, part-of-speech tagging

Suggested approach

- Motivation: discriminative classification methods are based on the assumption of the same distribution for training and testing data. However, this is not always true.
- Intuition: common representation can make two domains appear similar and enable effective domain adaptation.
- Formalization: a bound on the *target* generalization error of a classifier trained from labelled data in the *source* domain.

Problem Setup

- $\mathcal{X}$ is the instance set,$\{0,1\}$ is the label set, $\mathcal{Z}$ is a feature set, e.g. $\mathbb{R}^d$;
- distribution $\mathcal{D}$ over $\mathcal{X}$, a target function $f : \mathcal{X} \longrightarrow [0,1]$ – common for both domains ;
- a representation function $\mathcal{R} : \mathcal{X} \longrightarrow \mathcal{Z}$ and hypothesis class $\mathcal{H} \subseteq \{g : \mathcal{Z} \longrightarrow \{0,1\}\}$
- $\mathcal{R}$ induces distribution $\widetilde{\mathcal{D}}$ over $\mathcal{Z}$ and its subsets and function $\tilde{f} : \mathcal{Z} \longrightarrow [0,1]$:

$$\Pr_{\widetilde{\mathcal{D}}}[B] = \Pr_{\mathcal{D}}[\mathcal{R}^{-1}(B)]$$
$$\tilde{f}(z) = \mathrm{E}_{\mathcal{D}}[f(x)|\mathcal{R}(x) = z]$$

- For a predictor $h : \mathcal{Z} \longrightarrow [0,1]$ the expected error is $\epsilon_{\mathcal{T}}(h) = \mathrm{E}_{z \sim \widetilde{\mathcal{D}}_{\mathcal{T}}} \left| \tilde{f}(z) - h(z) \right|$

Assumptions

- There is a hypothesis $h \in \mathcal{H}$ that performs well on *both* domains, i.e. there is exists h and a small λ such that $\inf_{h \in \mathcal{H}}[\epsilon_{\mathcal{S}}(h) + \epsilon_{\mathcal{T}}(h)] \leq \lambda$

- $\mathcal{H}$ has bounded capacity, e.g. VC-dimension d.

Generalization Bound

Theorem

If a random labeled sample of size m is generated by applying $\mathcal{R}$ to a $\mathcal{D}_S$ - i.i.d. sample labeled according to f, and $\widetilde{\mathcal{U}}_S$, $\widetilde{\mathcal{U}}_T$ are unlabeled samples of size m′ each, drawn from $\widetilde{\mathcal{D}}_S$ and $\widetilde{\mathcal{D}}_T$ respectively, then with probability at least $1-\delta$ (over the choice of the samples), for every $h \in \mathcal{H}$:

$$\epsilon_T(h) \leq$$

$$\hat{\epsilon}_S(h) + \frac{4}{m}\sqrt{d \log \frac{2em}{d} + log\frac{4}{\delta}} + \lambda + d_{\mathcal{H}}(\widetilde{\mathcal{U}}_S, \widetilde{\mathcal{U}}_T) + 4\sqrt{\frac{d \log(2m\prime) + \log \frac{4}{\delta}}{m'}}$$

A good representation $\mathcal{R}$ achieves low values for training error and domain distance simaltaneously.

Computing the distance

- Distance between distributions $\mathcal{D}, \mathcal{D}'$ is defined as $d_{\mathcal{A}}(\mathcal{D}, \mathcal{D}') = 2\sup_{A \in \mathcal{A}} |\Pr_{\mathcal{D}}[A] - \Pr_{\mathcal{D}'}[A]|$,
- Domain distance is to be measure only with respect to function in the hypothesis class where $d_{\mathcal{H}}(.,.)$ indicates the distance on the class of subsets $\mathcal{Z}_h = \{z \in \mathcal{Z} : h(z) = 1\}$.
- A proxy for $d_{\mathcal{H}}(\widetilde{\mathcal{U}}_{\mathcal{S}}, \widetilde{\mathcal{U}}_{\mathcal{T}})$ is obtained by training a classifier to discriminate between points generated by *source* and *target* distributions.

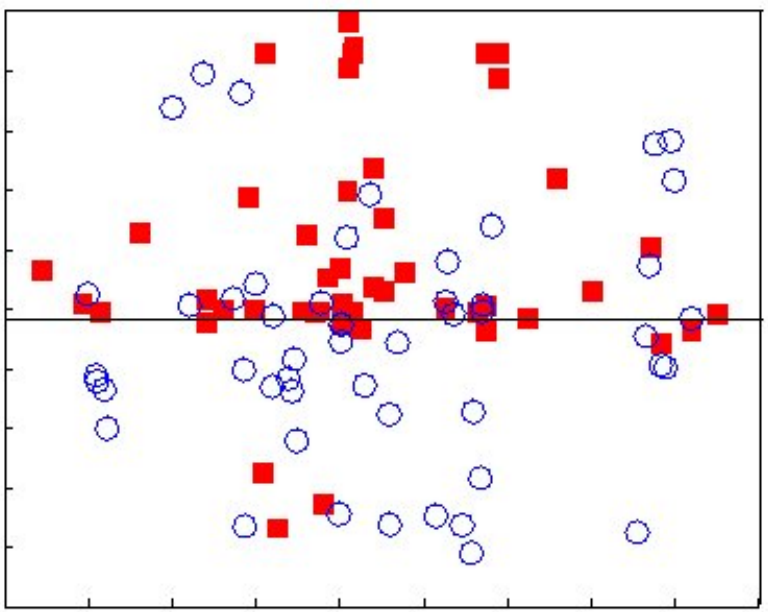
Adapting a part-of-speech tagger from the financial to biomedical domains

- Procedure:
  - choose a representation $\mathcal{R}$;
  - train a linear classifier using $\mathcal{R}$;
  - measure both relevant terms of the bound.
- Building $\mathcal{R}$ (by Structural Correspondence Learning):
  - find domain-independent "pivot" features occuring frequently in the unlabelled data in **both** domains, e.g., determiners, $<$ the token on the left $>$;
  - represent other features using their relative co-occurence counts with the pivot features;
  - compute a low-dimensional approximation to the co-occurrence matrix (through the singular value decomposition of the matrix).
- Intuition: features from *source* and *target* domains which behave similarly for PoS tagging will be represented similarly in the projected space.

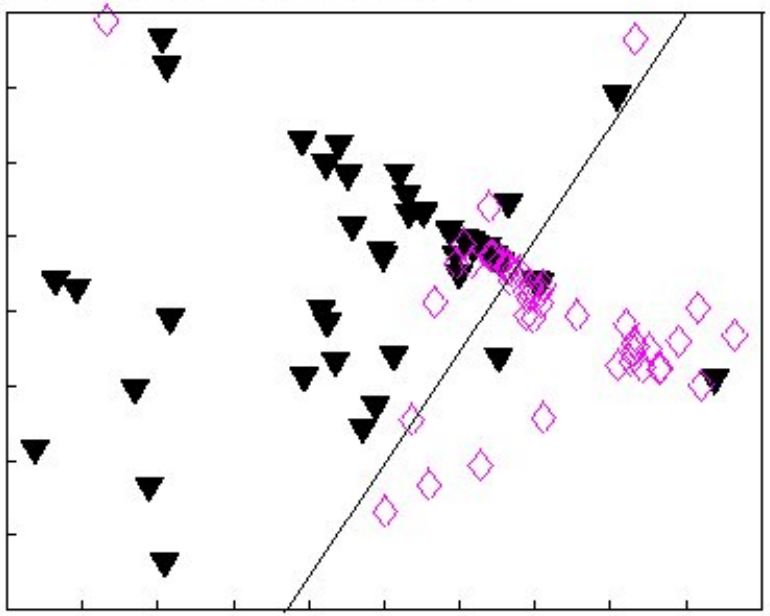
Empirical Results

- *Source* : articles from Wall Street Journal (WSJ)
- Labelled: 100 part-of-speech tagged sentences from WSJ
- *Target* data: biomedical abstracts (MEDLINE)
- Unlabelled data: 500,000 words from each domain
- instances are high-dimensional, binary vectors; $\mathcal{Z}$ is $\mathbb{R}^d$, $d = 200$;

(a) Plot of SCL representation for financial (squares) vs. biomedical (circles)



(b) Plot of SCL representation for nouns (diamonds) vs. verbs (triangles)



Contributions

Analysis of a classification problem when a *source* domain and a *target* domain have **different** distributions
An upper bound on the generalization of a classifier trained on a *source* domain and applied on a *target* domain

Important References

"Detecting Change in Data Streams", by D. Kifer, S. Ben-David, J. Gehrke, *Proc of the 30th VLDB Conference*, 2004.
"Domain Adaptation with Structural Correspondence Learning", by J. Blitzer, R. McDonald, F. Pereira, *Proc of EMNLP*, 2006.