Projection based transfer learning

Christian Poelitz
Dortmund Technical University

Transfer Learning

We want to reuse a trained model or information from different data sources to classify a new data set. We assume to have labelled data from data source S and want to learn a classifier on a unlabelled data source T. We use kernel methods in order leverage different high-dimensional features for a classification task.

Transfer Learning on Subspaces

We assume that the different data sources share similarities in low dimensional subspaces. These subspaces are invariant across the data sources and contain the information that are characteristic in both sources. Using only this information a classifier trained on source S might also perform well on source T.

Distances in Hilbert Spaces

We want to project onto a subspace such that the maximum mean discrepancy measure (Gretton et al. [GBR+08]) is minimized.

$$MMD(F,S,T) = sup_{f\in F}(\frac{1}{|S|}\sum_{x\in S} f(x) - \frac{1}{|T|}\sum_{x\in T} f(x)) >$$

$$MMD_P(F,S,T) = sup_{f\in P\circ F}(\frac{1}{|S|}\sum_{x\in S} f(x) - \frac{1}{|T|}\sum_{x\in T} f(x))$$

Subspace Methods

Kernel PCA: $K = n \cdot C = \sum_i \phi(x_i) \cdot \phi(x_i)^T$ for $\{x_i \in T \cup S\}$.
An eigenvalue decomposition on C results in a set of eigenvalues $\{\lambda_i\}$ and eigenvectors $\{v_i\}$ such that $\lambda_i \cdot v_i = C \cdot v_i$.
The projection onto the first k eigenvalues:
$P_U(\phi(x)) =$
$(\sum_j \alpha_{j,1} < \phi(x_i), \phi(x) >, \cdots, \sum_j \alpha_{j,k} < \phi(x_i), \phi(x) >)$
with $\alpha_{i,j} = (\frac{1}{\sqrt{\lambda_i}} \cdot v_i)_j$.

Other Subspace Methods

Subspace Alignment as proposed by Feranando et al. [FHST13] cannot be used since in kernel methods the projections must be in the sample (kernel defined sub) space. Hence, our projections must be expansions of the data samples. The cross kernel must be used to project all examples from both sources into the same Hilbert space. The approach by Zhang et al. [ZZW^{+}13] via surrogate kernels might be applicable and will be investigated in the future.

Efficiency

- Kernel methods scale quadratic or even cubic in the number of examples.
- We want to select only those examples that are close to the invariant subspace.
- This reduces the size of the kernel.

Greedy Selection

Distance based (Shawe Taylor et al. [STC04]):

$$x_{t+1} = argmin_{x \in S - \{x_1, \cdots, x_t\}} \| P_{U_T}(\phi(x)) \|^2$$

Herding based (Chen et al. [CWS12]):

$$x_{t+1} = argmax_{x \in S - \{x_1, \cdots, x_t\}} < w_t, \phi(x) >$$
$$w_{t+1} = w_t + E_{p_T}[\phi(x)] - \phi(x_{t+1})$$

Iteratively add examples and project all data onto the spanned subspace. If MMD between the different sources does increase rapidly, stop. This will be further investigated in the future.

Experiments

Method	E→D	E→B	E→K	D→E	D→B	D→K
kPCA	75.9	73.9	81.3	74	77.7	75
KMM	68.7	70.7	81.8	70.7	74.3	74.1
TCA	64.7	65.2	80.3	73.7	69.5	77.2
kPCA+	74,2	72.1	80.6	73.2	76	74.4
kPCAμ	74.9	68.4	81.2	70.6	76.2	72.5
Method	B→E	B→D	B→K	K→E	K→D	K→B
kPCA	71.9	77.5	72.7	84.4	79.8	76
KMM	68	71.2	69.6	83.9	73.5	74.6
TCA	73	69	73.8	76.7	67.8	63.7
kPCA+	71.7	75.1	70.2	82.9	79	76.5
kPCAμ	67.5	76.1	70.6	82.1	78	77.3

Table: This table shows the accuracies on target domains using training data from different source domains, *Source* → *Target*. Methods: Kernel Mean Matching (KMM), kernel PCA, Distance Based (kPCA+) and Kernel Herding Based (kPCAμ).

Experiments

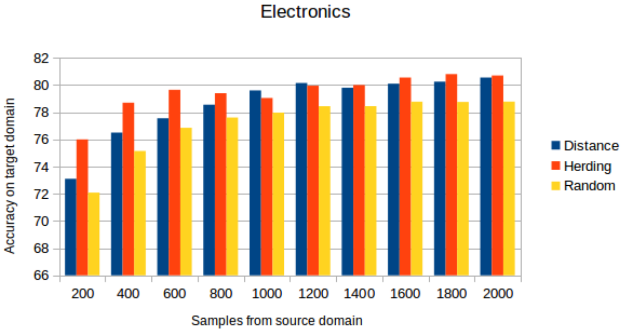

Figure: Results on the target data domain for the different categories. We compare random samples with our greedy selection strategy for sampling.

Issues tackled in the future

- Choose U_T and $U_{T \cup S'}$ w.r.t. distribution of the eigenvalues of K_T, resp. $K_{T \cup S'}$
- Investigate which kernels to use
- There are kernels for which $E_{p_T}[\phi(x)]$ cannot be efficiently computed
- Comparison to other (non-greedy) approaches (for instance Gong et al. [GGS13])
- Investigation on stopping criteria
- Further experiments including significance tests
- Convergence bounds

(Far) Future Work

- Extension to multi kernel settings

Questions?

Questions?

Thanks for your attention!

Yutian Chen, Max Welling, and Alex J. Smola.
Super-samples from kernel herding.
CoRR, abs/1203.3472, 2012.

Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars.
Unsupervised visual domain adaptation using subspace alignment.
In *ICCV*, 2013.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola.
A kernel method for the two-sample problem.
CoRR, abs/0805.2368, 2008.

Boqing Gong, Kristen Grauman, and Fei Sha.
Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation.
In *ICML (1)*, volume 28 of *JMLR Proceedings*, pages 222–230. JMLR.org, 2013.

John Shawe-Taylor and Nello Cristianini.
Kernel Methods for Pattern Analysis.
Cambridge University Press, New York, NY, USA, 2004.

Kai Zhang, Vincent Zheng, Qiaojun Wang, James Kwok, Qiang Yang, and Ivan Marsic.
Covariate shift in hilbert space: A solution via sorrogate kernels.
In Sanjoy Dasgupta and David Mcallester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 388–395. JMLR Workshop and Conference Proceedings, May 2013.