Regularized Semi-Paired Kernel CCA for Domain Adaptation[1]

Siamak Mehrkanoon Johan A.K. Suykens

KU Leuven, ESAT-STADIUS, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

Abstract

A Regularized Semi-Paired Kernel Canonical Correlation Analysis (RSP-KCCA) formulation is introduced for learning a latent space for the domain adaptation problem. The optimization problem is formulated in the primal-dual LS-SVM setting where side information can be readily incorporated through regularization terms. A joint representation of the data set across different domains is learned by solving a generalized eigenvalue problem or linear system of equations in the dual. The proposed model is naturally equipped with out-of-sample extension property which plays an important role for model selection. Experimental results are given to illustrate the effectiveness of the proposed approaches on synthetic and real-life datasets.

1 Introduction

Manual labeling of sufficient training data for diverse application domains is a costly, laborious task and often prohibitive. Therefore, designing models that can leverage rich labeled data in one domain and be applicable to a different but related domain is highly desirable. In particular, domain adaptation or transfer learning algorithms seek to generalize a model trained in a source domain (training data) to a new target domain (test data). Depending on the availability of the labeled instances in both domains, three scenarios can be considered, i.e. unsupervised, supervised and semi-supervised domain adaptation models [2]. Unsupervised domain adaptation approaches, do not take label information into consideration when learning the feature representation. On the other hand, supervised domain adaptation approaches, only use labeled data from the source and target domains. In the semi-supervised setting, one learns from labeled source instances as well as a small fraction of the target labeled instances. This setting can have many real world applications, as collecting labeled instances might be costly.

2 Formulation of the method

Consider two training datasets $\mathscr{D}^{(1)}$ and $\mathscr{D}^{(2)}$ as source and target domains with

$$\mathscr{D}^{(1)} = \{ \underbrace{x_1^{(1)},...,x_{n_p}^{(1)}}_{\substack{paired\\(labeled/unlabeled)\\(\mathscr{D}^{(1)}_{p,ul}\cup\mathscr{D}^{(1)}_{p,l})}} , \underbrace{x_{n_p+1}^{(1)},..,x_{n_1}^{(1)}}_{\substack{unpaired\\(unlabeled)\\(\mathscr{D}^{(1)}_{up,ul})}} , \underbrace{x_{n_p+1}^{(1)},..,x_{N_1}^{(1)}}_{\substack{unpaired\\(labeled)\\(\mathscr{D}^{(1)}_{up,l})}} \}, \quad \mathscr{D}^{(2)} = \{ \underbrace{x_1^{(2)},...,x_{n_p}^{(2)}}_{\substack{paired\\(labeled/unlabeled)\\(\mathscr{D}^{(2)}_{p,ul}\cup\mathscr{D}^{(2)}_{p,l})}} , \underbrace{x_{n_p+1}^{(2)},..,x_{n_2}^{(2)}}_{\substack{unpaired\\(unlabeled)\\(\mathscr{D}^{(2)}_{up,ul})}} \},$$

where $\mathscr{D}_p^{(1)}$ and $\mathscr{D}_p^{(2)}$ are paired samples whereas $\mathscr{D}_{up}^{(1)}$ and $\mathscr{D}_{up}^{(2)}$ are unpaired samples. Often in a domain adaptation setting, the number of labeled instances in the target domain is limited, compared to that of the source domain. Therefore, here we assume that only a small number of paired labeled data points from both domains are available, i.e. $\{x_1^{(i)},\ldots,x_{n_L}^{(i)}\}$ from $\mathscr{D}_p^{(1)}$ and $\mathscr{D}_p^{(2)}$ are labeled for $i = 1,2$. Furthermore, the source dataset is equipped with additional unpaired labeled instances during training. Assume that there are Q classes, then the label indicator matrix $Y^{(1)} \in \mathbb{R}^{n_{L_1}\times Q}$ for the source

[1]The full paper has been accepted for publication in *IEEE-TNNLS*, DOI: 10.1109/TNNLS.2017.2728719

domain is defined as $Y_{ij}^{(1)} = +1$, if the ith point belongs to the jth class and -1 otherwise. Similarly one can define the label indicator matrix $Y^{(2)} \in \mathbb{R}^{n_{L_2} \times Q}$ for the target domain where n_{L_2} denotes the total number of labeled instances in the target domain. The Regularized Semi-Paired KCCA (RSP-KCCA) with centered implicit feature map matrices $\Phi_c^{(1)}$ and $\Phi_c^{(2)}$ in the primal is formulated as follows [2]:

$$\begin{aligned} \max_{w_\ell^{(1)}, w_\ell^{(2)}, r_\ell, e_\ell} \quad & \mu \sum_{\ell=1}^{Q} e_\ell^T A r_\ell - \frac{1}{2} \sum_{i=1}^{2} \sum_{\ell=1}^{Q} {w_\ell^{(i)}}^T w_\ell^{(i)} - \frac{1}{2} \sum_{\ell=1}^{Q} {e_\ell}^T V_1 e_\ell - \frac{1}{2} \sum_{\ell=1}^{Q} {r_\ell}^T V_2 r_\ell + \frac{\gamma_3}{2} \sum_{\ell=1}^{Q} e_\ell^T c_\ell^{(1)} + r_\ell^T c_\ell^{(2)} \\ \text{subject to} \quad & e_\ell = \Phi_c^{(1)} w_\ell^{(1)},\ \ell = 1, \ldots, Q, \\ & r_\ell = \Phi_c^{(2)} w_\ell^{(2)},\ \ell = 1, \ldots, Q, \end{aligned}$$

where $c_\ell^{(1)}$ is the ℓ-th column of the matrix $C^{(1)}$ defined as $C^{(1)} = [c_1^{(1)}, \ldots, c_Q^{(1)}] = \left[\begin{array}{c} Y^{(1)} \\ \hline 0_{n_{u_1} \times Q} \end{array} \right]_{N_1 \times Q}$, and the subscript n_{u_1} denotes the total number of unlabeled instances from source domain. $0_{n_{u_1} \times Q}$ is a zero matrix of size $n_{u_1} \times Q$. Similarly $c_\ell^{(2)}$ is the ℓ-th column of the matrix $C^{(2)}$ defined as $C^{(2)} = [c_1^{(2)}, \ldots, c_Q^{(2)}] = \left[\begin{array}{c} Y^{(2)} \\ \hline 0_{n_{u_2} \times Q} \end{array} \right]_{n_2 \times Q}$, where the subscript n_{u_2} denotes the total number of unlabeled instances from target domain. Here $V_1 = \gamma_1 P_1 + (1-\gamma_1) L_1$ and $V_2 = \gamma_2 P_2 + (1-\gamma_2) L_2$. Matrix A, P_1 and P_2 are defined as follows: $A = \left[\begin{array}{c|c} I_{n_p \times n_p} & 0_{n_p \times n_2 - n_p} \\ \hline 0_{N_1 - n_p \times n_p} & 0_{N_1 - n_p \times n_2 - n_p} \end{array} \right], P_1 = \left[\begin{array}{c|c} I_{n_p \times n_p} & 0_{n_p \times N_1 - n_p} \\ \hline 0_{N_1 - n_p \times n_p} & 0_{N_1 - n_p \times N_1 - n_p} \end{array} \right], P_2 = \left[\begin{array}{c|c} I_{n_p \times n_p} & 0_{n_p \times n_2 - n_p} \\ \hline 0_{n_2 - n_p \times n_p} & 0_{n_2 - n_p \times n_2 - n_p} \end{array} \right]$ and here L_1 and L_2 are the graph Laplacian matrices associated with each of the dataset $\mathscr{D}^{(1)}$ and $\mathscr{D}^{(2)}$.

The applicability of the approach is shown on the Multiple Features Dataset taken from the UCI Machine Learning repository. It contains ten classes, $(0-9)$, of handwritten digits with 200 images per class, thus for a total of 2000 images. These digits are represented in terms of six different types of features (heterogeneous features) with different dimensionalities. We first select at random 50% from both source and target data in order to construct the paired training samples. Then 20% of the remaining data from both source and target domains is selected at random to be used as unpaired and unlabeled samples. The final remaining 30% of the source data are labeled in order to construct unpaired and labeled. The same remaning percentage of the target domain data defines the test target dataset, $\mathscr{D}_{\text{test}}^{(2)}$.

Table 1: Comparing the average test accuracy of the proposed RSP-KCCA model with those of mSDA model over 10 simulation runs.

	mSDA [1]						**RSP-KCCA** [2]					
Source domain	**fac**	**fou**	**kar**	**mor**	**pix**	**zer**	**fac**	**fou**	**kar**	**mor**	**pix**	**zer**
fac	0.9601	0.1840	0.2453	0.3000	0.6217	0.3743	0.9560	0.7683	0.8987	0.6510	0.9233	0.7190
fou	0.6597	0.7953	0.3350	0.3003	0.6123	0.4147	0.8573	0.7993	0.8520	0.6360	0.7853	0.7860
kar	0.6457	0.3170	0.9440	0.2810	0.5483	0.3487	0.9270	0.7780	0.9454	0.6583	0.9540	0.7947
mor	0.5410	0.4083	0.2420	0.6976	0.5167	0.3230	0.7503	0.6857	0.8290	0.7344	0.7920	0.7347
pix	0.6243	0.1733	0.3480	0.2313	0.9644	0.3823	0.9503	0.7817	0.9177	0.6253	0.9590	0.7773
zer	0.6920	0.2697	0.2333	0.2943	0.5770	0.8031	0.8400	0.7837	0.8040	0.6340	0.8773	0.8063

Acknowledgments. The research leading to these results received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC AdG A-DATADRIVE-B (290923). This letter reflects only our views: The EU is not responsible for any use that may be made of the information in it. The research leading to these results received funds from the following sources: Research Council KUL: GOA/10/09 MaNet, CoE PFV/10/002 (OPTEC), BIL12/11T; PhD/Postdoc grants; Flemish Government: FWO: PhD/Postdoc grants, projects: G.0377.12 (Structured systems), G.088114N (Tensor based data similarity); IWT: PhD/Postdoc grants, projects: SBO POM (100031); iMinds Medical Information Technologies SBO 2014; Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO, Dynamical systems, control and optimization, 2012-2017). Siamak Mehrkanoon is a postdoctoral fellow of the Research Foundation-Flanders (FWO).

References

[1] Minmin Chen, Zhixiang Xu, Kilian Weinberger, and Fei Sha. Marginalized denoising autoencoders for domain adaptation. In *in Proceedings of the 29th International Conference on International Conference on Machine Learning (ICML)*, pages 1627–1634, 2012.

[2] Siamak Mehrkanoon and Johan AK Suykens. Regularized semipaired kernel CCA for domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 2017,DOI: 10.1109/TNNLS.2017.2728719.