Sparse Domain Adaptation in a Good Similarity-Based Projection Space

Emilie Morvant, Amaury Habrard, Stéphane Ayache

Sparse Domain Adaptation in a Good Similarity-Based Projection Space*

Emilie Morvant, Stéphane Ayache
Aix-Marseille Univ, LIF-Qarma
CNRS, UMR 6166, 13013, Marseille, France
firstname.lastname@lif.univ-mrs.fr

Amaury Habrard
Univ of St-Etienne, Laboratoire Hubert Curien
CNRS, UMR 5516, 42000, St-Etienne, France
amaury.habrard@univ-st-etienne.fr

We address *domain adaptation* (DA) for binary classification in the challenging case where no target label is available. We propose an original approach that stands in a recent framework of Balcan *et al.* [1] allowing to learn linear classifiers in an explicit projection space based on *good similarity functions* that may be not symmetric and not positive semi-definite (PSD). Following the DA framework of Ben-David *et al.* [2], our method looks for a relevant projection space where the source and target distributions tend to be close. This objective is achieved by the use of an additional regularizer motivated by the notion of *algorithmic robustness* proposed by Xu and Mannor [3]. Our approach is formulated as a linear program with a 1-norm regularization leading to sparse models. We provide a theoretical analysis of this sparsity and a generalization bound. From a practical standpoint, to improve the efficiency of the method we propose an iterative version based on a reweighting scheme of the similarities to move closer the distributions in a new projection space. Hyperparameters and reweighting quality are controlled by a reverse validation process. The evaluation of our approach on a synthetic problem and real image annotation tasks shows good adaptation performances.

This work will appear in "IEEE International Conference on Data Mining (ICDM) 2011" [4].

Notations

Let $X \subseteq \mathbb{R}^d$ be the input space of dimension d and $Y = \{-1, +1\}$ the label set. A *domain* is a probability distribution over $X \times Y$. In a DA framework [2, 5], we have a *source domain* represented by a distribution P_S and a *target domain* represented by a somewhat different distribution P_T. D_S and D_T are the respective marginal distributions over X. A learning algorithm is provided with a *Labeled Source sample* $LS = \{(\mathbf{x}_i, y_i)\}_{i=1}^{d_l}$ drawn *i.i.d.* from P_S, and an *unlabeled Target Sample* $TS = \{\mathbf{x}_j\}_{j=1}^{d_t}$ drawn *i.i.d.* from D_T. Let $h : X \to Y$ be an hypothesis function. The expected source error of h over P_S is the probability that h commits an error: $err_S(h) = \mathbb{E}_{(\mathbf{x},y)\sim P_S} L_{01}\big(h,(\mathbf{x},y)\big)$, where $L_{01}(h,(\mathbf{x},y))=1$ if $h(\mathbf{x}) \neq y$ and zero otherwise, it is the 0-1 *loss function*. The target error err_T over P_T is defined in a similar way. $\hat{\mathrm{err}}_S$ and $\hat{\mathrm{err}}_T$ are the empirical errors. An hypothesis class $\mathcal{H}$ is a set of hypothesis. The DA objective is then to find a low target error hypothesis.

Domain Adaptation Framework

We consider the DA framework proposed by Ben-David *et al.* [2] allowing us to upper bound the target error err_T according to the source error and the divergence between the domain distributions,

$$\forall h \in \mathcal{H},\ \mathrm{err}_T(h) \ \leq\ \mathrm{err}_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \nu. \tag{1}$$

The last term ν can be seen as a kind of adaptation ability measure of $\mathcal{H}$ for the DA problem considered and corresponds to the error of the best joint hypothesis over the two domains:

*Work partially supported by the ANR VideoSense project (ANR-09-CORD-026) and the PASCAL2 network of excellence.

$\nu = \text{argmin}_{h\in\mathcal{H}} \text{err}_S(h) + \text{err}_T(h)$. The second term $d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T)$ is called the $\mathcal{H}\Delta\mathcal{H}$-distance between the two domain marginal distributions. This measure is actually related to $\mathcal{H}$ and an interesting point is that when the VC-dimension of $\mathcal{H}$ is finite, we can estimate $d_{\mathcal{H}\Delta\mathcal{H}}$ from finite samples by looking for the best classifier able to separate LS from TS. This bound suggests that one possible solution for a good DA is to look for a relevant data projection space where both the $\mathcal{H}\Delta\mathcal{H}$-distance and the source error of a classifier are low (two aspects *a priori* necessary for a good DA [6]).

Learning with Good Similarity Functions

Instead of working on the implicit high dimensional projection space induced by classical SVM's kernels (that may be strongly limited by symmetry and PSD requirements), we investigate a more flexible and intuitive similarity-based representation proposed recently by Balcan *et al.* [1] for learning with a **good similarity function** fulfilling the following definition.

Definition 1 ([1]). *A similarity function is any pairwise function $K : X \times X \to [-1, 1]$. K is an* **(ϵ,γ,τ)-good similarity function** *for a learning problem P if there exists a (random) indicator function $R(\mathbf{x})$ defining a set of reasonable points such that the following conditions hold:*
(i) A $1-\epsilon$ probability mass of examples $(\mathbf{x}, y)$ satisfy $\mathbb{E}_{(\mathbf{x}',y')\sim P}\big[yy'K(\mathbf{x},\mathbf{x}')|R(\mathbf{x}')=1\big] \geq \gamma$,
(ii) $Pr_{\mathbf{x}'}[R(\mathbf{x}')=1] \geq \tau$.

Def.1 requires that a large proportion of examples is on average more similar, *w.r.t* the margin γ, to the reasonable points of the same class than to the reasonable points of the opposite class. It includes all valid kernels as well as some non-PSD similarities and is thus a generalization of kernels [1].
Given K an (ϵ,γ,τ)-good similarity function, LS a sample of d_l labeled points and R a set of - enough - d_u potential reasonable points (*landmarks*), the conditions of Balcan *et al.* are sufficient to learn a low-error linear binary classifier (a *SF classifier*) in a ϕ^R-space defined by the mapping ϕ^R, which projects a point in the explicit space of the similarities to the landmarks in R such that,

$$\phi^R : \left\{ \begin{array}{ccc} X & \to & \mathbb{R}^{d_u} \\ \mathbf{x} & \mapsto & \langle K(\mathbf{x},\mathbf{x}'_1), \ldots, K(\mathbf{x},\mathbf{x}'_{d_u})\rangle. \end{array} \right.$$

The low-error SF classifier h can be learned by solving the following linear problem in the ϕ^R-space,

$$\min_{\boldsymbol{\alpha}} \frac{1}{d_l}\sum_{i=1}^{d_l} L\big(g,(\mathbf{x}_i,y_i)\big) + \lambda\|\boldsymbol{\alpha}\|_1, \text{ with } L\big(g,(\mathbf{x}_i,y_i)\big) = [1-y_i g(\mathbf{x})]_+ \text{ and } g(\mathbf{x}) = \sum_{j=1}^{d_u} \alpha_j K(\mathbf{x},\mathbf{x}'_j), \quad (2)$$

where $[1-z]_+ = \max(0, 1-z)$ is the hinge loss. Finally, we have $h(\mathbf{x}) = \text{sign}[g(\mathbf{x})]$.
Solving (2) not only minimizes the expected source error but also defines a relevant projection space for a given problem, since landmarks associated with a null weight in the solution $\boldsymbol{\alpha}$ will not be considered. In this work we propose to add a new regularization term on $\boldsymbol{\alpha}$ in order to constrain the explicit ϕ^R-space to move closer the two distributions and to tend to decrease the $\mathcal{H}\Delta\mathcal{H}$-distance.

Contribution for Domain Adaptation with Good Similarity Functions

The objective here is to define a regularizer that tends to make the source and target sample indistinguishable. For this purpose, we have investigated the *algorithmic robustness* notion proposed by Xu and Mannor [3] based on the following property: *"If a testing sample is similar to a training sample then the testing error is close to the training error"*. This can be formalized as follows: If for any test point close to a training point of the same class the deviation between the losses of each point is low for a learned model, then this model has some generalization guarantees (even if the robustness is true for only a subpart of the training sample). This result actually assumes that the test and train data are drawn from the same distribution and is thus not valid in a classical DA scenario.

However, we propose to adapt this principle for making the target sample similar to the source one which is coherent with the minimization of the divergence $d_{\mathcal{H}\Delta\mathcal{H}}$: For any pair $(\mathbf{x}_s,\mathbf{x}_t)$ of close source and target instances of the same class y, the deviation between the losses of $\mathbf{x}_s$ and $\mathbf{x}_t$ is low. By considering the hinge loss of (2), this leads us to the following term to minimize for such a pair: $\big|L(g,(\mathbf{x}_s,y)) - L(g,(\mathbf{x}_t,y))\big| \leq \big\|({}^t\phi^R(\mathbf{x}_s) - {}^t\phi^R(\mathbf{x}_t))\,\text{diag}(\boldsymbol{\alpha})\big\|_1$, where ${}^t\phi^R(\cdot)$ is the transposed vector of $\phi^R(\cdot)$ and $\text{diag}(\boldsymbol{\alpha})$ is the diagonal matrix with $\boldsymbol{\alpha}$ as main diagonal. Given any pair set

$\mathcal{C}_{ST}$ of close source-target examples, we then propose to consider this term for all the pairs as an additional regularizer on $\boldsymbol{\alpha}$, weighted by a parameter β, to the Problem (2) of Balcan *et al.* Our global optimization problem (3) can be then formulated as the following linear program,

$$\begin{cases} \min_{\boldsymbol{\alpha}} \frac{1}{d_l} \sum_{i=1}^{d_l} L\big(g, (\mathbf{x}_i, y_i)\big) + \lambda \|\boldsymbol{\alpha}\|_1 + \beta \sum_{(\mathbf{x}_s, \mathbf{x}_t) \in \mathcal{C}_{ST}} \|({}^t\phi^R(\mathbf{x}_s) - {}^t\phi^R(\mathbf{x}_t)) \operatorname{diag}(\boldsymbol{\alpha})\|_1, \\ \text{with } L\big(g, (\mathbf{x}_i, y_i)\big) = [1 - y_i g(\mathbf{x})]_+ \text{ and } g(\mathbf{x}) = \sum_{j=1}^{d_u} \alpha_j K(\mathbf{x}, \mathbf{x}'_j). \end{cases} \tag{3}$$

This problem is defined with a 1-norm regularization leading generally to very sparse models. We have in fact proved in the following lemma that the sparsity of the obtained models depends also on a quantity $B_R = \min_{\mathbf{x}'_j \in R} \left\{ \max_{(\mathbf{x}_s, \mathbf{x}_t) \in \mathcal{C}_{ST}} |K(\mathbf{x}_s, \mathbf{x}'_j) - K(\mathbf{x}_t, \mathbf{x}'_j)| \right\}$ related to the deviation between coordinates in the considered ϕ^R-space. In other words, when the domains are far from each other, *i.e.* the task is hard, B_R tends to be high which can increase the sparsity.

Lemma 1. *For any $\lambda > 0$, $\beta > 0$, any set $\mathcal{C}_{ST}$ s.t. $B_R > 0$, if $\boldsymbol{\alpha}^*$ is optimal, then $\|\boldsymbol{\alpha}^*\|_1 \leq \frac{1}{\beta B_R + \lambda}$.*

Moreover, according to the robustness framework [3] applied on the source domain, and from the DA bound (1), we can prove the following generalization bound for the expected target domain error.

Theorem 1. *Problem* (3) *defines a procedure* $(\mathbf{2M}_\eta, \frac{\mathbf{N}_\eta}{\beta \mathbf{B_R} + \lambda})$ **robust on the source domain**, *where* $N_\eta = \max_{\substack{\mathbf{x}_a, \mathbf{x}_b \sim D_S, \\ \rho(\mathbf{x}_a, \mathbf{x}_b) \leq \eta}} \|{}^t\phi^R(\mathbf{x}_a) - {}^t\phi^R(\mathbf{x}_b)\|_\infty$, *$\eta > 0$, M_η is the η-covering number of X. Thus for every h in the SF classifiers hypothesis class, for any $\delta > 0$, with probability at least $1 - \delta$,*

$$\operatorname{err}_T(h) \leq \hat{\operatorname{err}}_S(h) + \frac{N_\eta}{\beta B_R + \lambda} + \sqrt{\frac{4M_\eta \ln 2 + 2 \ln \frac{1}{\delta}}{d_l}} + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \nu.$$

From a practical standpoint, a critical issue is the estimation of the hyperparameters λ and β and the definition of the pair set $\mathcal{C}_{ST}$ which is difficult *a priori* since we do not have any target label information. We propose to tackle these problems with the help of a reverse validation method.

Reverse validation and Iterative procedure

We choose the different parameters of our method by following the principle of *reverse validation* of Zhong *et al.* [7]. This principle is illustrated on Fig. 1 and consists in learning a so-called *reverse classifier* h^r, from the target sample sef-labeled by a classifier h (inferred with Pb. (3)). We evaluate $\hat{\operatorname{err}}_S(h^r)$ on the source sample and heuristically $\hat{\operatorname{err}}_T(h^r)$ on the self-labeled target sample (both by cross-validation). We then obtain an heuristic estimate of ν (of the DA bound (1)) for h^r such that $\hat{\nu} = \hat{\operatorname{err}}_S(h^r) + \hat{\operatorname{err}}_T(h^r)$. We select the parameters and pair set $\mathcal{C}_{ST}$ minimizing $\hat{\nu}$. In this context, $\hat{\nu}$ can be seen as a quality measure of the ϕ^R-space found: If the two domains are sufficiently close and related then the reverse classifier should perform well on the source domain ([8]).

However, considering all the possible pairs for $\mathcal{C}_{ST}$ remains clearly intractable. In practice, we select only a limited number of examples for building $\mathcal{C}_{ST}$. We compensate the possible loss of information by an heuristic iterative procedure still allowing to move closer the two distributions. Suppose that at a given iteration l, with a similarity K_l, we obtain new weights $\boldsymbol{\alpha}^l$ by solving (3). Our regularization term can actually be seen as a $L1$-distance in a new ϕ^R_{l+1}-space: $\|({}^t\phi^R_l(\mathbf{x}_s) - {}^t\phi^R_l(\mathbf{x}_t)) \operatorname{diag}(\boldsymbol{\alpha})\|_1 = \|({}^t\phi^R_{l+1}(\mathbf{x}_s) - {}^t\phi^R_{l+1}(\mathbf{x}_t))\|_1$. ϕ^R_{l+1} corresponding to the mapping defined by the similarity K_{l+1} obtained from K_l by a conditional reweighting to each landmark: $\forall \mathbf{x}'_j \in R, K_{l+1}(\mathbf{x}, \mathbf{x}'_j) = \alpha^l_j K_l(\mathbf{x}, \mathbf{x}'_j)$ (K_{l+1} do not need to be PSD nor symmetric according to Def.1). We then iterate the process in the new ϕ^R_{l+1}-space and we stop at iteration l when $\hat{\nu}_{l+1}$ has reached a convergence point or has increased. The main steps of the recursive approach are summarized on Algorithm 1.

Experimental evaluation

Our method DASF has been evaluated on a toy problem and on real image annotation tasks and compared with SF method of Pb (2) and SVM with no adaptation, the semi-supervised Transductive-SVM (T-SVM) [9] and the iterative DA algorithm DASVM [8]. We used a Gaussian kernel for the

Algorithm 1 DASF (Domain Adaptation with Similarity Functions)

Input: similarity function K, set R, samples LS and TS

Output: classifier h_{DASF}

$h_0(\cdot) \leftarrow \text{sign}\left[\frac{1}{|R|}\sum_{j=1}^{|R|} K(\cdot, \mathbf{x}'_j)\right]$; $K_1 \leftarrow K$; $l \leftarrow 1$;

while The stopping criterion is not verified **do**

$\boldsymbol{\alpha}^l \leftarrow$ Solve Pb. (3) with K_l, $\mathcal{C}_{ST}$ and hyperparameters being selected by reverse validation;

$K_{l+1} \leftarrow$ Update K_l according to $\boldsymbol{\alpha}^l$; Update R; $l++$;

end while

return $h_{DASF}(\cdot) = \text{sign}\left[\sum_{x'_j \in R} \alpha^l_j K_l(\cdot, \mathbf{x}'_j)\right]$;

last three methods and a re-normalization of this kernel as a non PSD similarity function for SF and DASF (see [10]). All the averaged results show that DASF provide better and sparser models in general. As an illustration, Tab.1 gives results for a real image annotation task, where the source images are extracted from the PascalVOC'07 corpus and the target ones from the TrecVid'07 video corpus. Moreover, the iterative procedure always tend to decrease the distribution divergence with the iterations [4]. Among all the possible perspectives, we notably aim to consider some few target labels to help the search of a relevant projection space.

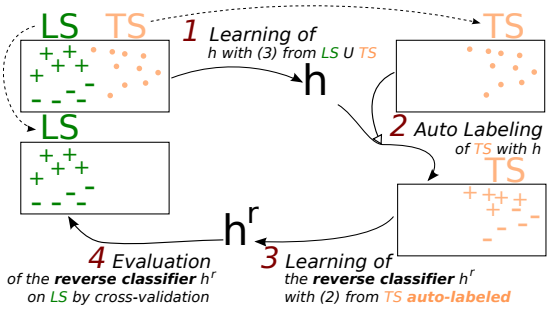

Figure 1: The reverse validation. **1**: *Learning* h with (3). **2**: *Auto-labeling* the target sample with h. **3**: *Learning* h^r on the auto-labeled target sample with (2). **4**: *Evaluation* of h^r on LS.

CONCEPT	BOAT	BUS	CAR	MONITOR	PERSON	PLANE	AVG.
SVM	0.56	0.25	0.43	0.19	0.52	0.32	0.38
MODEL SIZE	351	476	1096	698	951	428	667
SF	0.49	0.46	0.50	0.34	0.45	0.54	0.46
MODEL SIZE	214	224	**176**	246	226	178	211
T-SVM	0.56	0.48	0.52	0.37	0.46	0.61	0.50
MODEL SIZE	498	535	631	741	1024	259	615
DASVM	0.52	0.46	**0.55**	0.30	0.54	0.52	0.48
MODEL SIZE	202	222	627	523	274	450	383
DASF	**0.57**	**0.49**	**0.55**	**0.42**	**0.57**	**0.66**	**0.54**
MODEL SIZE	**120**	**130**	254	**151**	**19**	**7**	**113**

Table 1: The results obtained on the TrecVid target domains according to the F-measure. AVG. corresponds to the averaged results.

References

[1] M.-F. Balcan, A. Blum, and N. Srebro. Improved guarantees for learning via similarity functions. In *Proceedings of COLT*, 2008.

[2] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J.W. Vaughan. A theory of learning from different domains. *Machine Learning Journal*, 79(1-2):151–175, 2010.

[3] H. Xu and S. Mannor. Robustness and generalization. In *Proceedings of COLT*, 2010.

[4] E. Morvant, A. Habrard, and S. Ayache. Sparse domain adaptation in projection spaces based on good similarity functions. In *Proceedings of ICDM*, 2011.

[5] Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Proceedings of COLT*, 2009.

[6] S. Ben-David, T. Lu, T. Luu, and D. Pal. Impossibility theorems for domain adaptation. *JMLR W&CP*, 9:129–136, 2010.

[7] E. Zhong, W. Fan, Q. Yang, O. Verscheure, and J. Ren. Cross validation framework to choose amongst models and datasets for transfer learning. In *Proceedings of ECML-PKDD*, 2010.

[8] L. Bruzzone and M. Marconcini. Domain adaptation problems: A DASVM classification technique and a circular validation strategy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(5), 2010.

[9] T. Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of ICML*, 1999.

[10] E. Morvant, A. Habrard, and S. Ayache. On the usefulness of similarity based projection spaces for transfer learning. In *Proceedings of Similarity-Based Pattern Recognition workshop (SIMBAD)*, 2011.