Agnostic Domain Adaptation

Alexander Vezhnevets Joachim M. Buhmann

ETH Zurich
8092 Zurich, Switzerland
{alexander.vezhnevets,jbuhmann}@inf.ethz.ch

Abstract. The supervised learning paradigm assumes in general that both training and test data are sampled from the same distribution. When this assumption is violated, we are in the setting of transfer learning or domain adaptation: Here, training data from a *source* domain, aim to learn a classifier which performs well on a *target* domain governed by a different distribution. We pursue an agnostic approach, assuming no information about the shift between source and target distributions but relying exclusively on unlabeled data from the target domain. Previous works [2] suggest that feature representations, which are invariant to domain change, increases generalization. Extending these ideas, we prove a generalization bound for domain adaptation that identifies the transfer mechanism: what matters is how much learnt classier itself is invariant, while feature representations may vary. Our bound is much tighter for rich hypothesis classes, which may only *contain* invariant classifier, but can not be invariant altogether. This concept is exemplified by the computer vision tasks of semantic segmentation and image categorization. Domain shift is simulated by introducing some common imaging distortions, such as gamma transform and color temperature shift. Our experiments on a public benchmark dataset confirm that using domain adapted classifier significantly improves accuracy when distribution changes are present.

1 Introduction

The fundamental assumption in supervised learning is that training and test data arise from the same distribution. However in real life applications, it is common that training examples from one *source* domain are used to build the predictor that is expected to perform a related task on a different *target* domain. This change requires domain adaptation.

The situation with domain changes and the need for domain adaptation is shared by many fields, e.g., we have to account for domain change when we train a spam filter for a new user on examples from other users. In natural language processing, this occurs in e.g., part-of-speech tagging [5], where the tagger is trained on medical texts and deployed on legal texts. In computer vision systems, classifiers are usually trained prior to deployment on data which are manually annotated by experts. This labeling process is tedious and expensive, whereas data collection is usually fast and inexpensive. For instance, collecting unlabeled

data from a video surveillance system under different settings (different camera, lighting) requires little effort, but labeling this data demands a human annotator and often even a trained domain expert. Hence, the ability to adapt to a new domain using only unlabeled data from a new target distribution is of substantial practical advantage.

To achieve good generalization in supervised learning one should keep the hypothesis class simple while minimizing the empirical risk. Intuitively, in case of domain adaptation an additional requirement should be imposed: the source and the target distributions should look the same for the good classifier. In other words, classifier should be invariant to the change of the distribution.

We consider the setting where a finite set of *labeled* training examples is available from the *source* domain, but only few *unlabeled* examples are available from the *target* domain. We proceed by first proving a bound on the *target* generalization error, which is dependant on the classifier's training error on the source distribution and its invariance to distribution changes. This bound is much tighter for rich hypothesis classes, that only *contain* invariant hypothesis, but are quite variable in general. Invariance is formulated as the inability of the classifier to discriminate between the source and the target domain. Finally, we construct an algorithm that minimizes this bound.

Along with a theoretical analysis of domain adaptation we present an experimental validation of our results. Computer vision serves as a challenging application domain for machine learning, which allows us to visually inspect our results. We experimentally show the applicability of our approach by constructing a domain adaptive version of semantic texton forest [13] (STF) for image semantic segmentation and categorization. Semantic segmentation simultaneously requires to segment and recognize objects, one of the fundamental and most challenging computer vision tasks. We study the adaptation of STF to color cast and gamma transform, which are very natural distortions in imaging. We will see that such distortions are very damaging for an STF. But with domain adaptation we are able to improve results in some cases by more than a factor of two.

The paper is organized as follows. We first shortly describe previous works. Section 3 formally defines our problem and notation. In section 4 we present our theoretical bound. In section 5 the domain adaptive random forest for semantic segmentation and categorization is discussed. Section 6 describes our experimental results followed by a conclusion.

2 Prior Work

In this paper we consider the setting of domain adaptation for an *arbitrary shift* in the data distribution and where only *unlabeled* data from the target domain is available. Much research has been done to address each constraint individually. However little has been reported when both constraints are considered. We briefly review the literature for different settings of learning under distribution shift.

Transfer learning Transfer learning is a setting when labeled data from the target distribution is available. In [4] authors prove uniform convergence bounds

for algorithms that minimize a convex combination of source and target empirical risk. A thorough experimental evaluation for this scenario (minimization of convex combination of risks) can be found in [12]. In [8] a boosting method for transfer learning is developed. [11] studies adaptation with multiple sources, where for each source domain, the distribution over the input points as well as a hypothesis with error at most ϵ are given. They prove that combinations of hypotheses weighted by the source distributions benefit from favorable theoretical guarantees.

Covariate shift One common assumption to address the case where labels are not available is that of covariate shift. Here, only the marginal distribution $Pr[X]$ changes and the conditional remains unchanged, i.e., $Pr_{D_S}[Y|X] = Pr_{D_T}[Y|X]$. In [3] the general problem of learning under covariate shift is formulated as an integrated optimization problem and a kernel logistic regression classifier is derived for solving it. A nonparametric method which directly produces resampling weights without distribution estimation for learning under covariate shift is presented in [10]. Their method works by matching distributions between training and testing sets in feature space. Another paper [3] studies a complex problem of learning multiple tasks (multitask learning), when each task may have a covariate shift. They derive a learning procedure that produces resampling weights which match the pool of all examples to the target distribution of any given task.

Semi-supervised learning Semi-supervised learning (SSL) [6] is another strategy to improve the classifiers accuracy by using unlabeled data. Our setting should not be confused with semi-supervised learning. As in SSL we use unlabeled data to improve our classifier. While SSL assumes data to come from the *same* distribution, our setting does not impose such an assumption. One common approach to semi-supervised learning is to treat labels of unlabeled data as additional variables which have to be optimized to maximize the possible separation margin. Such strategy could also be advocated for domain adaptation as a valid heuristic (and was used for that purpose in [1]), though it has no theoretical support. Examples of semi-supervised methods are tSVM [14] and semi-supervised random forests [7].

Domain adaptation The setting that is closest to ours has been defined in [2]. The authors also consider the case with no assumptions about shift and they require only unlabeled data from target distribution. By studying the influence of feature representation on domain adaptation, they theoretically prove that the hypothesis space that is invariant to distribution changes improves generalization, although the problem of finding such a space is not addressed. In contrast to [2] we are interested in learning a classifier from a rich, possibly not invariant family, which generalizes well under distribution shift. We discuss this work in more details in Section 4.

3 Problem Setup

Let X be the instance set and Y be the set of labels. The joint distributions are given by $\tilde{D}_S(X \times Y)$ and $\tilde{D}_T(X \times Y)$, for the source and the target domains

respectively. The corresponding marginal distributions of X are denoted by D_S and D_T. To simplify the notation we restrict ourselves to dichotomies, i.e., to two classes. Labeled training samples are drawn from $\tilde{D}_S(X \times Y)$, but only samples of unlabeled data are gathered from D_T. Let $H \subseteq \{h : X \to Y\}$ be the hypothesis space. The probability that hypothesis h makes an error on the source domain as

$$\epsilon_S(h) = E_{(x,y)\sim\tilde{D}_S}[h(x) \neq y]. \tag{1}$$

The error on the target domain $\epsilon_T(h)$ is defined similarly. Z_h denotes the characteristic function of h,

$$Z_h = \{x \in X : h(x) = 1\}. \tag{2}$$

The symmetric set difference is abbreviated by $A\Delta B = (A \setminus B) \cup (B \setminus A)$. For example, $Pr_{D_S}[Z_h \Delta Z_{h^*}]$ is the probability that $[h \neq h^*]$ with respect to the marginal distribution of X in the source domain.

We do not make any assumptions about the nature of the domain shift. It is possible that both marginals $Pr(X)$ and conditional probabilities $Pr(X|Y)$ are changing and the resulting bound is completely agnostic.

4 Generalization Bound

Now we derive our theoretical results. Suppose there is a hypothesis in H which performs λ well on both domains:

$$\inf_{h\in H}[\epsilon_S(h) + \epsilon_T(h)] \leq \lambda. \tag{3}$$

In the work of Shai Ben-David [2], the following generalization bound was provided in a form dependant on the $\mathcal{A}$-distance between the source and the target domain:

$$\begin{aligned}\epsilon_T(h) \leq & \hat{\epsilon}_S(h) + \sqrt{\frac{4}{m}(d\log\frac{2em}{d} + \log\frac{4}{\delta})} \\ & + \lambda + d_{\mathcal{H}}(\mathcal{D}_{\mathcal{S}}, \mathcal{D}'_{\mathcal{T}}).\end{aligned} \tag{4}$$

In words, the $\mathcal{A}$-distance is proportional to an ability of family of predictors to distinguish between two distributions:

$$d_{\mathcal{A}}(\mathcal{D}, \mathcal{D}') = 2 \sup_{A\in\mathcal{A}} |Pr_{\mathcal{D}}[A] - Pr_{\mathcal{D}'}[A]|. \tag{5}$$

Unfortunately the question of how to find such a family of predictors was not addressed. The bound in eq. 4 states that when the features and the hypothesis family are invariant to domain shift, then we can expect to generalize well. However, the bound does not tell us how to choose the best hypothesis from the hypothesis class. Even the experimental results in [2] are not fully justified by this bound, since the feature representation was learnt *after* seeing the data. Formally, to apply this bound, all possible variants of feature representation, which

could be learnt by structural correspondence learning [5], should be included into hypothesis family H, which will render the bound trivial.

We will now show a bound, which depends on the characteristics of the particular hypothesis chosen by the training algorithm, rather than on the hypothesis class that we choose from. Our bound would allow us to design an algorithm, that explicitly searches for a hypothesis that minimizes it.

We formalize *invariance* of a classifier as its inability to distinguish between the source and the target distributions:

$$\psi(h) = |Pr_{D_S}[Z_h] - Pr_{D_T}[Z_h]|. \tag{6}$$

$\psi(h)$ has a minimum at zero (high invariance), i.e., the classifier cannot distinguish between the source and the target distribution. It exhibits a maximum of one (low invariance) when the classifier can always accurately decide from which distribution a data point comes from.

Now we are ready to formulate our theorem.

Theorem 1. *Let H be a hypothesis space of VC-dimension d. Given m i.i.d. samples from $\tilde{D}_S$, $\forall h \in H$, with probability of at least $1-\delta$,*

$$\begin{aligned}\epsilon_T(h) \leq & \hat{\epsilon}_S(h) + \sqrt{\frac{4}{m}(d\log\frac{2em}{d} + \log\frac{4}{\delta})} + \lambda \\ & + \psi(h) + \psi(h^*) + 2\psi(h\cdot h^*),\end{aligned} \tag{7}$$

where $\psi(h) = |Pr_{D_S}[Z_h] - Pr_{D_T}[Z_h]|$.

Proof. Let $h^* = \arg\min_{h\in H}(\epsilon_T(h) + \epsilon_S(h))$, and let λ_S and λ_T be the errors of h^* on the source and the target domains respectively. Note that $\lambda = \lambda_S + \lambda_T$.

To get the bound in [2] authors bounded the change, induced by the domain shift, of the difference between the best hypothesis h^* and the learnt h by the invariance of the hypothesis family to distribution shift – $\mathcal{A}$-distance. Essentially, we decompose this invariance into three parts: invariance of the learnt hypothesis, of the best hypothesis, and of their intersection. The proof is the following:

$$\epsilon_T(h) \leq \lambda_T + Pr_{D_T}[Z_h \Delta Z_{h^*}] \tag{8}$$

$$= \lambda_T + Pr_{D_S}[Z_h \Delta Z_{h^*}] - Pr_{D_S}[Z_h \Delta Z_{h^*}] + Pr_{D_T}[Z_h \Delta Z_{h^*}] \tag{9}$$

$$\begin{aligned}= & \lambda_T + Pr_{D_S}[Z_h \Delta Z_{h^*}] - Pr_{D_S}[Z_h \setminus Z_{h^*}] - Pr_{D_S}[Z_h^* \setminus Z_h] \\ & + Pr_{D_T}[Z_h \setminus Z_{h^*}] + Pr_{D_T}[Z_h^* \setminus Z_h]\end{aligned} \tag{10}$$

$$\begin{aligned}= & \lambda_T + Pr_{D_S}[Z_h \Delta Z_{h^*}] + Pr_{D_T}[Z_h] - Pr_{D_T}[Z_h \cap Z_{h^*}] - Pr_{D_S}[Z_h] \\ & + Pr_{D_S}[Z_h \cap Z_{h^*}] + Pr_{D_T}[Z_{h^*}] - Pr_{D_T}[Z_h \cap Z_{h^*}] \\ & - Pr_{D_S}[Z_{h^*}] + Pr_{D_S}[Z_h \cap Z_{h^*}]\end{aligned} \tag{11}$$

$$\begin{aligned}\leq & \lambda_T + Pr_{D_S}[Z_h \Delta Z_{h^*}] + |Pr_{D_T}[Z_h] - Pr_{D_S}[Z_h]| \\ & + |Pr_{D_T}[Z_{h^*}] - Pr_{D_S}[Z_{h^*}]| \\ & + 2\,|Pr_{D_T}[Z_h \cap Z_{h^*}] - Pr_{D_S}[Z_h \cap Z_{h^*}]|\end{aligned} \tag{12}$$

$$= \lambda_T + Pr_{D_S}[Z_h \Delta Z_{h^*}] + \psi(h) + \psi(h^*) + 2\psi(h\cdot h^*) \tag{13}$$

$$\leq \underbrace{\lambda + \psi(h^*)}_{\text{constant}} + \underbrace{\epsilon_S(h) + \psi(h) + 2\psi(h\cdot h^*)}_{\text{dependant on } h}. \tag{14}$$

To finish the proof one needs to apply classic Vapnik-Chervonenkis [14] theory to bound $\epsilon_s(h)$ by its empirical estimate. Using Vapnik-Chervonenkis theory again, we can bound the true ψ by its empirical estimate and an additional complexity penalty.

Observe that the bound has two parts: a constant part and the second part that is dependant on h. The constant (w.r.t. h) part is a function of the hypothesis class, which is assumed to be fixed. The part that depends on h consists of a sum of the classifier's training error $\epsilon_S(h)$ and the invariances $\psi(h)$ and $\psi(h \cdot h^*)$. The term $\psi(h)$ is the invariance of a learnt classifier, which can be controlled during training. The term $\psi(h \cdot h^*)$ measures the intersection between the learnt h and the optimal classifier h^*. It is large when the overlap between h and h^* changes a lot after the domain shift. Regretfully $\psi(h \cdot h^*)$ can neither be measured nor optimized, since we completely lack and knowledge of h^*: it pinpoints the uncertainty incurred in the absence of labeled data in the target domain. Hence in our design of the algorithm in section 5 we will only minimize the empirical estimates of $\epsilon(h)$ and $\psi(h)$.

In contrast to eq. 4 [2], we no longer rely on the invariance of the entire hypothesis class, but rather only on the specific classifier that we learn. Thus optimization of the bound can be integrated into the training process directly, as demonstrated in the following section. Our bound is also tighter for rich hypothesis classes, where invariant hypothesis is contained, but the class itself far from being invariant.

5 Algorithm

Theorem 1 provides us with an insight on how such a domain adaptive classifier could be constructed. When minimizing empirical loss on the source distribution, $\psi(h)$ should also be minimized by increasing its invariance. We implement this idea for the random forest classifier. In particular, for extremely randomized forests [9] (ERF), a predicate for splitting is selected that concurrently maximizes information gain and minimizes the empirical estimate of $\psi(h)$. Here we describe an ERF for the computer vision task of semantic segmentation – a task of simultaneous object segmentation and recognition. This ERF will also provide us with an adapted kernel for SVM based object categorization.

Semantic Texton Forest The Semantic Texton Forest (STF) proposed in [13] is employed for semantic segmentation. Their work uses ERF for pixel-wise classification. Below we shortly describe this approach.

A decision forest is an ensemble of K decision trees. A decision tree works by recursively branching left or right down the tree according to a learnt binary split function $\phi_n(x) : X \to \{0, 1\}$ at the node n, until a leaf node l is reached. Associated with each leaf l in the tree is a learned class distribution $P(c|l)$. Classification is done by averaging over the leaf nodes $L = (l_1, ..., l_K)$ reached for all K trees:

$$P(c|L) = \frac{1}{K} \sum_{k=1}^{K} P(c|l_k). \tag{15}$$

Conceptually, a forest consists of a structure, consisting of nodes with split functions, and probability estimates in the leaf nodes. We can represent a forest as a complex function $f(g(x))$, where $g : X \to N^K$ maps the instance feature vectors to the vector of leaf indices, reached for each tree and $f : N^K \to [0,1]^C$ maps those indices to class probability estimates. Each leaf then has to store a vector $w^l = [P(y = 1|l), ..., P(y = C|l)]$.

Trees are trained independently on random subsets of the training data. Learning proceeds recursively, splitting the training data at node into left and right subsets according to a split function $\phi_n(x)$. At each split node, several candidates for $\phi_n(x)$ are generated randomly, and the one that maximizes the expected gain in information about the node categories is chosen:

$$\Delta H = -\frac{|I_l|}{|I_n|}H(I_l) - \frac{|I_r|}{|I_n|}H(I_r), \tag{16}$$

where $H(I)$ is the Shannon entropy of the classes in the set of examples I. The recursive training usually continues to a fixed maximum depth without pruning. The class distributions $P(y|l)$ are estimated empirically as a histogram of the class labels y_i of the training examples i that reached leaf l.

STF, as presented above, provides prediction on the basis of local context only. To bring a global image context into play an Image Level Priors (ILP) are used. The support vector machine (SVM) is trained to predict whether a certain object is present in the image. Output of SVM is scaled to the probability simplex and pixel level STF predictions are then multiplied by it. A kernel for the SVM is constructed by matching the amount of pixels in two given images that pass through the same nodes in the STF. We refer the reader to the original publication [13] for more details on ILP and STF training.

Domain Adaptation To adapt a STF to a particular domain, we introduce a slight modification to the original criterion (eq 16) for choosing the best split function $\phi_n(x)$. The new criterion $\Delta\tilde{H}$ now takes shift invariance into account as desired

$$\Delta\tilde{H} = \underbrace{\Delta H^S}_{\widehat{\epsilon(h)}} - \alpha \underbrace{\left(\left| \frac{|I_r^T|}{|I_n^T|} - \frac{|I_r^S|}{|I_n^S|} \right| + \left| \frac{|I_l^T|}{|I_n^T|} - \frac{|I_l^S|}{|I_n^S|} \right| \right)}_{\widehat{\psi(h)}}, \tag{17}$$

where I_r^T and I_l^T are the data points from the target domain right and left of the split respectively. I_n^T is the total amount of the target domain data points that have been classified by a node. The same notation is used for the source data with respective superscripts $I_{(\cdot)}^S$. First term – ΔH^S stands for information gain on labeled data from source domain and optimizes an empirical estimate of $\epsilon(h)$. The second term is an empirical estimate of $\psi(h)$. This addendum deals with unlabeled data from both domains, penalizing those splits, that produce classifiers invariant, which are not invariant to the distribution shift. This modification, forces our classifier to both minimize error on the source distribution and maximize invariance of the classifier towards distribution changes. It slightly

increases the training time and has no effect on the computational complexity of the final predictor. Adaptation of the image categorizer emerges in a natural way, since the adapted STF provides the (adapted) kernel for the image categorizer. The proposed approach is generic and can possibly be applied to other application fields too.

6 Experiments

We evaluate our approach on two fundamental computer vision tasks: semantic segmentation and image categorization. The benefit of using visual data for domain adaptation experiments is that we can introduce realistic distribution shifts based on common imaging distortions and visually inspect the results. For our experiments, we used the MSRC21 dataset. This dataset comprises of 591 images out of 21 object classes. The standard train/test/validation split, as in [13], contains 276/256/59 images, respectively. In order to estimate standard error deviation we used 5 random splits of the dataset into train/test/validation sets keeping the same proportions as in the standard split. We applied several common imaging distortions on the dataset to simulate distribution shift. We train our classifiers on the undistorted training images. The unlabeled set of distorted validation images is used for adaptation.

We compare our algorithm (**STF-DA**) with two baseline methods. The first baseline is a **STF** [13] trained on the undistorted training set only. We also compared our results to a semi-supervised random forest (**STF-SSL**) [7] trained on the training set and unlabeled validation set, for the following reasons. First, it can be readily integrated into the STF framework. Second, it optimizes the separation margin of the classifier over all classifier parameters and all labelings of unlabeled data, which is a valid heuristic for domain adaptation in case when no information is available on the distribution shift and labeled data for target domain are lacking. We evaluate on distorted test images. For all classifiers, we use the implementation with the parameter setting of the STF framework as provided in [13].

Image Distortions
We consider two distortions common in imaging: color temperature shift and gamma transform. Distortions are applied to test and validation set. Both distortion types change both the *marginal* and the *conditional* probabilities. Color shift affects only color features, when gamma transform inflicts a nonlinear change in all features. Moreover, the hypothesis class (random forests) are far from being invariant towards this distortions. We introduce two versions of the shift for both distortions. One is deterministic - every image is perturbed in the same way (shift parameters are constant). In the second case, for each image we randomly select a distortion parameter. In contrast to the previous works we are able to deal with this setting both in theory and in practice.

Color Temperature Shift Color temperature is a characteristic of visible light that has important applications in lighting and photography. The color temperature of a light source is determined by comparing its chromaticity with that of an ideal black-body radiator. Color temperature shift is a common artifact

Table 1. Accuracy on semantic segmentation and image categorization tasks.

Distortion	Semantic Segmentation			Image Categorization		
	STF	**STF-SSL**	**STF-DA**	**STF**	**STF-SSL**	**STF-DA**
Temp. (det)	0.19 ± 0.02	0.20 ± 0.03	0.44 ± 0.05	0.25 ± 0.02	0.26 ± 0.02	0.48 ± 0.04
Temp. (rand)	0.37 ± 0.03	0.38 ± 0.02	0.46 ± 0.03	0.40 ± 0.02	0.40 ± 0.01	0.52 ± 0.03
Gamma (det)	0.48 ± 0.04	0.50 ± 0.03	0.52 ± 0.03	0.53 ± 0.04	0.53 ± 0.04	0.58 ± 0.02
Gamma (rand)	0.41 ± 0.03	0.42 ± 0.02	0.45 ± 0.03	0.44 ± 0.02	0.45 ± 0.03	0.49 ± 0.02

of digital photography. The same scene shot under different lighting will have a color cast: the warm yellow-orange cast of tungsten lamps or the blue-white of florescent tubes. Most digital cameras perform white balance correction by digitally adjusting color temperature. For the deterministic case we reduced the temperature of all images by 40%. In the randomized case images have there temperature lowered by $40, 30, 20\%$ or 10% at random. Deterministic shift is very strong and renders nearly all color feature non reliable.

Gamma Transform Due to a finite dynamic range and discretization in digital cameras, images can easily become over- or under-exposed. We mimic this effect by the gamma transform $\tilde{p}_{i,j,c} = p_{i,j,c}^{\gamma}$. In our experiments we use $\gamma = 2$ for the deterministic case. Again, we have also produced a dataset with γ being randomly chosen in the interval $[0, 4]$ to make the shift non deterministic. This shift does not change images as dramatically as color shift, but is not restricted to a certain feature subspace. One can not adapt to it by just simply discarding certain features (as it could be done in color shift case).

Results

We evaluate on semantic segmentation task measuring overall per pixel accuracy and on the task of image categorization measuring average precision (Table 1).

In all experiments, **STF-DA** outperforms both baseline algorithms. **STF-SSL** fails to bring any significant improvement over **STF**: semi-supervised learning is inappropriate to account for a distribution shift. The most significant improvements of **STF-DA** over the baselines are observed on the data with color temperature shift. Our algorithm is able to filter out unreliable color features and perform better – in the deterministic case the accuracy increases *more then twice*. Success on γ transformed data validates that our approach is applicable even when the shift affects all features and when it is not restricted to only a subset of features. The more general non deterministic shifts are also processed satisfactorily by **STF-DA**.

7 Conclusion

We have presented an analysis of domain adaptation for cases where only unlabeled examples from the target distribution are available and no assumptions are made about the shift between the target and the source distributions. Intuitively, a good classifier should be invariant to changes in the distribution. We

formalize this intuition by proving an upper bound of the generalization error of classifiers trained on the *source* domain and tested on the *target* domain. Our bound explicitly depends on classifier's invariance and its error on the source distribution. In contrast to previous work [2] that requires the whole hypothesis class to be invariant, this study demonstrates that good generalization *can* be achieved even when the hypothesis family only *contains* one invariant classifier. We experimentally confirm our findings on the challenging tasks of semantic segmentation and image categorization. We show that our adaptation algorithm significantly improves results for different imaging distortions, in some cases by more than twice.

Acknowledgements

This work has been supported by the Swiss National Science Foundation under grant #200021-117946.

References

1. Arnold, A., Nallapati, R., Cohen, W.W.: A comparative study of methods for transductive transfer learning. In: In ICDM Workshop on Mining and Management of Biological Data (2007)
2. Ben-David, S., Blitzer, J., Crammer, K., Pereira, F.: Analysis of representations for domain adaptation. In: NIPS (2007)
3. Bickel, S., Brückner, M., Scheffer, T.: Discriminative learning for differing training and test distributions. In: ICML. ACM Press (2007)
4. Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Wortman, J.: Learning bounds for domain adaptation. In: NIPS (2007)
5. Blitzer, J., Mcdonald, R., Pereira, F.: Domain adaptation with structural correspondence learning. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Sydney, Australia (2006)
6. Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi-Supervised Learning. MIT Press, Cambridge, MA (2006)
7. Christian Leistner, Amir Saffari, J.S.H.B.: Semi-supervised random forests. In: ICCV (2009)
8. Dai, W., Yang, Q., Xue, G.R., Yu, Y.: Boosting for transfer learning. In: ICML. New York, NY, USA (2007)
9. F. Moosmann, B.T., Jurie., F.: Fast discriminative visual codebooks using randomized clustering forests. In: NIPS (2006)
10. Huang, J., Smola, A.J., Gretton, A., Borgwardt, K.M., Sch?olkopf, B.: Correcting sample selection bias by unlabeled data. In: NIPS (2006)
11. Mansour, Y., Mohri, M., Rostamizadeh, A.: Domain adaptation with multiple sources. In: NIPS (2009)
12. Schweikert, G., Widmer, C., Scho"lkopf, B., Ra"tsch, G.: An empirical analysis of domain adaptation algorithms for genomic sequence analysis. In: NIPS (2008)
13. Shotton, J., Johnson, M., Cipolla., R.: Semantic texton forests for image categorization and segmentation. In: ECCV (2008)
14. Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience (September 1998)