Transductive Adaptation of Black Box Predictions

Stéphane Clinchant, Gabriela Csurka and Boris Chidlovskii
Xerox Research Centre Europe
6 chemin Maupertuis, Meylan, France
Firstname.Lastname@xrce.xerox.com

Abstract

Access to data is critical to any machine learning component aimed at training an accurate predictive model. In reality, data is often a subject of technical and legal constraints. Data may contain sensitive topics and data owners are often reluctant to share them. Instead of access to data, they make available decision making procedures to enable predictions on new data. Under the *black box classifier constraint*, we build an effective domain adaptation technique which adapts classifier predictions in a transductive setting. We run experiments on text categorization datasets and show that significant gains can be achieved, especially in the unsupervised case where no labels are available in the target domain.

1 Introduction

While huge volumes of unlabeled data are generated and made available in various domains, the cost of acquiring data labels remains high. Domain Adaptation problems arise each time when one leverage labeled data in one or more related *source* domains, to learn a classifier for unseen data in a *target* domain which is related, but not identical. The majority of domain adaptation methods makes an assumption of *largely available* source collections; this allows to measure the discrepancy between distributions and either build representations common to both target and sources, or directly reuse source instances for a better target classification (Xu and Sun, 2012).

Numerous approaches have been proposed to address domain adaptation for statistical machine translation (Koehn and Schroeder, 2007), opinion mining, part of speech tagging and document ranking (Daumé, 2009), (Pan and Yang, 2010), (Zhou and Chang, 2014). Most effective techniques include feature replication (Daumé, 2009), pivot features (Blitzer et al., 2006), (Pan et al., 2010) and finding topic models shared by source and target collections (Chen and Liu, 2014). Domain adaptation has equally received a lot of attention in computer vision (Gopalan et al., 2015) where *domain shift* is a consequence of changing conditions, such as background, location and pose, etc.

More recently, domain adaptation has been tackled with word embedding techniques or deep learning. (Bollegala et al., 2015) proposed an unsupervised method for learning domain-specific word embedding while (Yang and Eisenstein, 2014) relied on *word2vec* models (Mikolov et al., 2013) to compute feature embedding. Deep learning has been considered as a generic solution to domain adaptation (Vincent et al., 2008; Glorot et al., 2011), (Chopra et al., 2013) and transfer learning problems (Long et al., 2015). For instance, denoising autoencoders are successful models which find common features between source and target collection. They are trained to reconstruct input data from partial random corruption and can be stacked into a multi-layered network where the weights are fine-tuned with backpropagation (Vincent et al., 2008) or marginalized out (Chen et al., 2012).

Domain adaptation is also very attractive for service companies operating customer business processes as it can reduce annotation costs. For instance, opinion mining components deployed in a service solution can be customized to a new customer and adapted with few annotations in order to achieve a contractual performance.

But, *in reality*, the simplifying assumption of having access to source data rarely holds and limits therefore the application of existing domain

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 326–331,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

adaptation methods. Source data are often a subject of legal, technical and contractual constraints between data owners and data customers. Often, customers are reluctant to share their data. Instead, they often put in place *decision making procedures*. This allows to obtain predictions for new data under *a black box scenario*. Note that this scenario is different from the differential privacy setting (Dwork and Roth, 2014) in the sense that no queries to the raw source database are allowed whereas, in our case, only requests for predicting labels of target documents are permitted. This makes privacy preserving machine learning methods inapplicable here (Chaudhuri and Monteleoni, 2008), (Agrawal and Srikant, 2000).

In addition, black boxes systems are frequent in natural language processing applications. For instance, Statistical Machine Translation (SMT) systems are often used as black box to extract features (Specia et al., 2009). Similarly, the problem of adapting SMT systems for cross lingual retrieval has been addressed in (Nikoulina et al., 2012) where target document collections cannot be accessed and the retrieval engine works as a black box.

In this paper we address the problem of adapting classifiers trained on the source data and available as *black boxes*. The case of available source classifiers has been studied by (Duan et al., 2009) to regularize supervised target classifiers, but we consider here a transductive setting, where the source classifiers are used to predict class scores for a set of available target instances.

We then apply the denoising principle (Vincent et al., 2008) and consider these predictions on target instances as corrupted by the domain shift from the source to target. More precisely, we use the stacked Marginalized Denoising Autoencoders (Chen et al., 2012) to *reconstruct* the predictions by exploiting the correlation between the target features and the predicted scores. This method has the advantage of coping with *unsupervised cases* where no labels in the target domain is available. We test the prediction denoising method on two benchmark text classification datasets and demonstrate its capacity to significantly improve the classification accuracy.

2 Transductive Prediction Adaptation

The domain adaptation problem consists of leveraging the source labeled and target unlabeled data to derive a hypothesis performing well on the target domain. To achieve this goal, most DA methods compute correlation between features in source and target domains. With no access to source data, we argue that the above principle can be extended to *the correlation between target features and the source class decisions*. We tune an *adaptation trick* by considering predicted class scores as augmented features for target data. In other words, we use the source classifiers *as a pivot* to transfer knowledge from source to target. In addition, one can exploit relations between the predictions scores and the target feature distribution to provide adapted predictions.

2.1 Marginalized Denoising Autoencoder

The *stacked Marginalized Denoising Autoencoder* (sMDA) is a version of the multi-layer neural network trained to reconstruct input data from partial random corruption (Vincent et al., 2008) proposed by (Chen et al., 2012), where the random corruption is marginalized out yielding the optimal reconstruction weights in the closed form.

The basic building block of the method is a one-layer linear denoising autoencoder where a set of N input documents $\mathbf{x}_n$ are corrupted M times by random feature dropout with the probability p. It is then reconstructed with a linear mapping $\mathbf{W} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ by minimizing the squared reconstruction loss[1]:

$$\mathcal{L}(\mathbf{W}) = \sum_{n=1}^{N} \sum_{m=1}^{M} ||\mathbf{x}_n - \mathbf{W}\tilde{\mathbf{x}}_{nm}||^2. \tag{1}$$

Let $\bar{\mathbf{X}}$ be the concatenation of M replicated version of the original data and $\tilde{\mathbf{X}}$ be the matrix representation of the M corrupted versions.

Then, the solution of (1) can be expressed as the closed-form solution for ordinary least squares $\mathbf{W} = \mathbf{P}\mathbf{Q}^{-1}$ with $\mathbf{Q} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top$ and $P = \bar{\mathbf{X}}\tilde{\mathbf{X}}^\top$, where the solution depends on the re-sampling of $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and which features are randomly corrupted.

It is preferable to consider all possible corruptions of all possible inputs when the denoising transformation $\mathbf{W}$ is computed, i.e. letting $m \rightarrow \infty$. By the weak law of large numbers, the matrices $\mathbf{P}$ and $\mathbf{Q}$ converge to their expected values $\mathbb{E}[\mathbf{Q}], \mathbb{E}[\mathbf{P}]$ as more copies of the corrupted data

[1] A constant is added to the input, $\mathbf{x}_n = [\mathbf{x}_n; 1]$, and an appropriate bias, never corrupted, is incorporated within $\mathbf{W}$.

are created. In the limit, one can derive their expectations and express the corresponding mapping for $\mathbf{W}$ in a closed form as $\mathbf{W} = \mathbb{E}[\mathbf{P}]\,\mathbb{E}[\mathbf{Q}]^{-\mathbf{1}}$, where:

$$\mathbb{E}[\mathbf{Q}]_{ij} = \left[\begin{array}{lll} \mathbf{S}_{ij} q_i q_j, & \text{if} & i \neq j, \\ \mathbf{S}_{ij} q_i, & \text{if} & i = j, \end{array} \right.$$

and $\mathbb{E}[\mathbf{P}]_{ij} = \mathbf{S}_{ij} q_j$ where $q = [1-p, \ldots, 1-p, 1] \in \mathbb{R}^{d+1}$ and $\mathbf{S} = \mathbf{X}\mathbf{X}^\top$ is the covariance matrix of the uncorrupted data. This closed form denoising layer with a unique noise p is referred in the following as *marginalized denoising autoencoder* (MDA).

It was shown by (Chen et al., 2012) that MDA can be applied with success to domain adaptation where the source set $\mathbf{X_s}$ and target set $\mathbf{X_t}$ are concatenated to form $\mathbf{X}$ and the mapping $\mathbf{W}$ can exploit the correlation between source and target features. The case of fully available source and target data is referred as a *dream case* in the evaluation section.

2.2 Prediction Adaptation

Without access to $\mathbf{X}_s$, MDA cannot be directly applied to $[\mathbf{X}_s; \mathbf{X}_t]$. Instead, we augment the feature set $\mathbf{X}_t$ with the class predictions represented as vector $f^s(\mathbf{x}^t)$ of class predictions $P_s(Y = y|\mathbf{x}_n^t), n = 1, \ldots, N$. Let $\mathbf{u}_n^t = [\mathbf{x}_n^t; f^s(\mathbf{x}_n^t)]$ be the target instance augmented with the source classifier predictions and $\mathbf{U} = [\mathbf{u}_1^t \mathbf{u}_2^t \ldots \mathbf{u}_N^t]$ be the input to the MDA. Then we compute the optimal mapping $\mathbf{W}^* = \min_{\mathbf{W}} ||\mathbf{U} - \mathbf{W}\tilde{\mathbf{U}}||^{\mathbf{2}}$ that takes into account the correlation between the target features $\mathbf{x}^t$ and class predictions $f^s(\mathbf{x}^t)$. The reconstructed class predictions can be obtained as $\mathbf{W}^*_{[1:N,d+1:d+C]} \cdot f^s(\mathbf{x}^t)$, where C is the number of classes, and used to label the target data. Algorithm 1 summarizes all steps of the transductive prediction adaptation for a single source domain; the generalization to multiple sources is straightforward[2].

Algorithm 1 Transductive prediction adaptation.

```
Require: Unlabeled target dataset $\mathbf{X}_t \in \mathbb{R}^{N \times d}$.
Require: Class predictions $f^s(\mathbf{x}^t) = [P_s(Y = 1|\mathbf{x}_i^t), \ldots, P_s(Y = C|\mathbf{x}_n^t)] \in \mathbb{R}^C$.
 1: Compose $\mathbf{U} \in \mathbb{R}^{N \times (d+C)}$ with $\mathbf{u}_n^t = [\mathbf{x}_n^t; f^s(\mathbf{x}_n^t)]$.
 2: Use MDA with noise level $p$ to estimate $\mathbf{W}^* = \min_{\mathbf{W}} ||\mathbf{U} - \mathbf{W}\tilde{\mathbf{U}}||^{\mathbf{2}}$.
 3: Get the denoised class predictions for $\mathbf{x}^t$ as $\mathbf{y}^t = \mathbf{W}^*_{[1:N,d+1:d+C]} \cdot f^s(\mathbf{x}^t)$.
 4: Label $\mathbf{x}^t$ with $c^* = \mathrm{argmax}_c\{y_c^t|\mathbf{y}^t\}$.
 5: return Labels for $\mathbf{X}_t$.
```

3 Experimental results

We test our approach on two standard domain adaptation datasets: the Amazon reviews (AMT) and the 20Newsgroups (NG). The AMT dataset consists of products reviews with 2 classes (positive and negative) represented by tf-idf normalized bag-of-words, used in previous studies on domain adaptation (Blitzer et al., 2011). We consider the 10,000 most frequent features and four domains used in the studies: *kitchen* (k), *dvd* (d), *books* (b) and *electronics* (e) with roughly 5,000 documents per domain. We use all the source dataset as training and test on the whole target dataset. We set the MDA noise level p to high values (*e.g.* 0.9), as document representations are sparse and adding low noise have no effect on the features already equal to zero.

In Table 1, we show the performance of the Transductive Prediction Adaptation (TPA) on 12 adaptation tasks in the AMT dataset. The first column shows the accuracies for *the dream case* where the standard MDA is applied to both source and target data. The second column shows the baseline results ($f^s(\mathbf{X}^t)$) obtained directly as class predictions by the source classifier. The classification model is an l_2 regularized Logistic Regression[3] cross-validated with regularized parameter $C \in [0.0001, 0.001, 0.1, 1, 10, 50, 100]$.

The two last columns show the results obtained with two versions of TPA (results are underlined when improving over the baseline and in bold when yielding the highest values). In the first version, target instances $\mathbf{x}_n^t$ contains only features (words and bigrams) appearing in the source documents and used to make the predictions $f(\mathbf{x}_n^t)$. In the second version, denoted as TPAe, we extend TPA with words unseen in the source documents. If the extension part is denoted $\mathbf{v}_n^t$, we obtain an augmented representation $\mathbf{u}_n^t = [\mathbf{x}_n^t; \mathbf{v}_n^t; f(\mathbf{x}_n^t)]$ as input to MDA.

[2] It requires concatenating the class predictions from different sources at step 1 and averaging the reconstructed predictions per class at step 3.

[3] We also experimented with other classifiers, such as SVM , Multinomial Naive Bayes, and obtained similar improvement after applying TPA. Results are not shown due to the space limitation.

Table 1: TPA results on the AMT dataset.

$S \rightarrow T$	**MDA***	$f^s(\mathbf{X}^t)$	**TPA**	**TPAe**
$d \rightarrow b$	*84.59*	81.36	82.61	**83.19**
$e \rightarrow b$	*78.07*	73.87	75.93	**79.95**
$k \rightarrow b$	*78.75*	73.50	75.02	**78.39**
$b \rightarrow d$	*85.07*	82.54	83.56	**84.32**
$e \rightarrow d$	*79.99*	76.46	77.67	**81.60**
$k \rightarrow d$	*80.76*	77.58	79.16	**81.92**
$b \rightarrow e$	*80.32*	76.44	78.54	**81.81**
$d \rightarrow e$	*83.70*	78.65	80.75	**82.89**
$k \rightarrow e$	*89.05*	87.55	88.38	**88.50**
$b \rightarrow k$	*84.00*	79.46	81.44	**85.21**
$d \rightarrow k$	*86.08*	80.83	83.15	**86.14**
$e \rightarrow k$	*90.76*	89.97	**91.10**	90.86
Avg	*83.4*	79.85	81.44	**83.73**

As we can see, both TPA and TPAe significantly outperform the baseline $f^s(\mathbf{X}^t)$ obtained with no adaptation. Furthermore, extending TPA with words present in target documents only allows to further improve the classification accuracy in most cases. Finally, TPAe often outperforms the *dream case* and also on average (note however that MDA* uses the features common to source and target documents as input).

To understand the effect of prediction adaptation we analyze the *book* $\rightarrow$ $electronics$ adaptation task. In the mapping $\mathbf{W}$, we sort the weights corresponding to the correlation between the positive class and the target features. Features with the highest weights (up-weighted by TPA) are *great, my, sound, easy, excellent, good, easy_to, best, yo, a_great, when, well, the_best*. On contrary, the words that got the smallest weight (down-weighted by TPA) are *no, was, number, don't, after, money, if, work, bad, get, buy*.

As TPA is totally unsupervised, we run additional experiments to understand its practical usefulness. We compare TPA to the case of *weakly annotated* target data, where few target examples are labelled and used for training a target classifier. Trained with 40, 100 and 200 target examples, a logistic regression yields an average accuracy of 64.63%, 68.01% and 75.13% over 12 tasks and a Multinomial Naives Bayes reports 65.82%, 71.49% and 76%, respectively. Even with 200 labeled target documents, the target versus target classification results are significantly below the 79.8% average accuracy of the baseline source classifier.

All these values are therefore significantly below the 83.73% obtained with TPAe. This strongly supports the domain adaptation scenario, when a sentiment analysis classifier trained on a larger source set and adapted to target documents can do better than a classifier trained on a small set of labeled target documents. Furthermore, we have seen that the baseline can be significantly improved by TPA and even more by TPAe without the need of even a small amount of manual labeling of the target set.

The second group of evaluation tests is on the 20Newsgroup dataset. It contains around 20,000 documents of 20 classes and represents a standard testbed for text categorization. For the domain adaptation, we follow the setting described in (Pan et al., 2012). We filter out rare words (appearing less than 3 times) and keep at most 10,000 features for each task with a tf-idf termweighting. As all documents are organized as a hierarchy, the domain adaptation tasks are defined on category pairs with sources and targets corresponding to subcategories. For example, for the *'comp vs sci'* task, subcategories such as *comp.sys.ibm.pc.hardware* and *sci.crypt* are set as source domains and *comp.sys.ibm.mac.hardware* and *sci.med* as targets, respectively.

In our experiments we consider 5 adaptation tasks on category pairs (*'comp vs sci'*,*'rec vs talk'*, *'rec vs sci'*, *'sci vs talk'* and *'comp vs rec'* as in (Pan et al., 2012)), and run the baseline, TPA and TPAe methods. For each category pair, we additionally inverse the source and target roles; this explains two sets of experimental results for each pair. We show the evaluation results in Table 2. It is easy to observe again the significant improvement over the baseline $f^s(\mathbf{x}_n^t)$ and the positive effect of including the unseen words in the TPA.

Table 2: TPA results on the 20Newsgroup dataset.

class pair	$f^s(\mathbf{X}^t)$	**TPA**	**TPAe**
'comp vs sci'	71.06	80.24	**80.43**
	65.4	71.6	**71.98**
'rec vs talk'	65.66	68.01	**70.18**
	69.93	75.84	**77.2**
'rec vs sci'	76.02	85.97	**86.42**
	74.17	81.14	**82.71**
'sci vs talk'	76.1	80.22	**81.3**
	74.92	80.07	**80.19**
'comp vs rec'	86.63	91.56	**92.06**
	86.97	92.67	**93.34**
Avg	74.69	80.73	**81.58**

4 Conclusion

In this paper we address the domain adaptation scenario without access to source data and where source classifiers are available as *black boxes*. In the transductive setting, the source classifiers can predict class scores for target instances, and we consider these predictions as corrupted by domain shift. We use the Marginalized Denoising Autoencoders (Chen et al., 2012) to reconstruct the predictions by exploiting the ”correlation” between the target features and the predicted scores. We test the transductive prediction adaptation on two known benchmarks and demonstrate that it can significantly improve the classification accuracy, comparing to the baseline and to the case of full access to source data. This is an encouraging result because it demonstrates that domain adaptation can still be effective despite the absence of source data. Lastly, in the future, we would like to explore the adaptation of other language processing components, such as named entity recognition, with our method.

References

[Agrawal and Srikant2000] Rakesh Agrawal and Ramakrishnan Srikant. 2000. Privacy-preserving data mining. In *ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pages 439–450.

[Blitzer et al.2006] John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 120–128.

[Blitzer et al.2011] John Blitzer, Sham Kakade, and Dean P. Foster. 2011. Domain adaptation with coupled subspaces. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 173–181.

[Bollegala et al.2015] Danushka Bollegala, Takanori Maehara, and Ken-ichi Kawarabayashi. 2015. Unsupervised cross-domain word representation learning. In *Annual Meeting of the Association for Computational Linguistics(ACL)*, pages 730–740.

[Chaudhuri and Monteleoni2008] Kamalika Chaudhuri and Claire Monteleoni. 2008. Privacy-preserving logistic regression. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 289–296.

[Chen and Liu2014] Zhiyuan Chen and Bing Liu. 2014. Topic modeling using topics from many domains, lifelong learning and big data. In *International Conference on Machine Learning (ICML)*, pages 703–711.

[Chen et al.2012] Minmin Chen, Zhixiang Xu, Kilian Q. Weinberger, and Fei Sha. 2012. Marginalized denoising autoencoders for domain adaptation. In *International Conference on Machine Learning (ICML)*, pages 767–774.

[Chopra et al.2013] Sumit Chopra, Suhrid Balakrishnan, and Raghuraman Gopalan. 2013. DLID: Deep learning for domain adaptation by interpolating between domains. In *ICML Workshop on Challenges in Representation Learning (WREPL)*.

[Daumé2009] H. Daumé. 2009. Frustratingly easy domain adaptation. *CoRR*, arXiv:0907.1815.

[Duan et al.2009] Lixin Duan, Ivor W. Tsang, Dong Xu, and Tat-Seng Chua. 2009. Domain adaptation from multiple sources via auxiliary classifiers. In *International Conference on Machine Learning (ICML)*, pages 289–296.

[Dwork and Roth2014] Cynthia Dwork and Aaron Roth. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9:211–407.

[Glorot et al.2011] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *International Conference on Machine Learning (ICML)*, pages 513–520.

[Gopalan et al.2015] Raghuraman Gopalan, Ruonan Li, Vishal M. Patel, and Rama Chellappa. 2015. Domain adaptation for visual recognition. *Foundations and Trends in Computer Graphics and Vision*, 8(4).

[Koehn and Schroeder2007] Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *ACL Workshop on Statistical Machine Translation (STAT-MT)*, pages 224–227.

[Long et al.2015] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. 2015. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning (ICML)*.

[Mikolov et al.2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, arXiv:1301.3781.

[Nikoulina et al.2012] Vassilina Nikoulina, Bogomil Kovachev, Nikolaos Lagos, and Christof Monz. 2012. Adaptation of statistical machine translation model for cross-lingual information retrieval in a service context. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 109–119.

[Pan and Yang2010] Sinno J. Pan and Qiang Yang. 2010. A survey on transfer learning. *Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

[Pan et al.2010] Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In *International Conference on World Wide Web (WWW)*.

[Pan et al.2012] Weike Pan, Erheng Zhong, and Yang Qiang. 2012. Transfer learning for text mining. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 223–257. Springer.

[Specia et al.2009] Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello N. Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *Annual Conference of the European Association for Machine Translation (EAMT)*, pages 28–35.

[Vincent et al.2008] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning (ICML)*.

[Xu and Sun2012] Zhijie Xu and Shiliang Sun. 2012. Multi-source transfer learning with multi-view adaboost. In *Annual Conference on Neural Information Processing Systems (NIPS)*, volume LNCS 7665, pages 332–339. Springer.

[Yang and Eisenstein2014] Yi Yang and Jacob Eisenstein. 2014. Unsupervised multi-domain adaptation with feature embeddings. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, pages 672–682.

[Zhou and Chang2014] Mianwei Zhou and Kevin C. Chang. 2014. Unifying learning to rank and domain adaptation: Enabling cross-task document scoring. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD*, pages 781–790.