# Domain Adaptation meets Active Learning

**Piyush Rai, Avishek Saha, Hal Daumé III, and Suresh Venkatasubramanian**
School of Computing, University of Utah
Salt Lake City, UT 84112
{piyush,avishek,hal,suresh}@cs.utah.edu

## Abstract

In this work, we show how active learning in some (target) domain can leverage information from a different but related (source) domain. We present an algorithm that harnesses the source domain data to learn the best possible initializer hypothesis for doing active learning in the target domain, resulting in improved label complexity. We also present a variant of this algorithm which additionally uses the domain divergence information to selectively query the most informative points in the target domain, leading to further reductions in label complexity. Experimental results on a variety of datasets establish the efficacy of the proposed methods.

## 1 Introduction

Acquiring labeled data to train supervised learning models can be difficult or expensive in many problem domains. Active Learning tries to circumvent this difficultly by only querying the labels of the most informative examples and, in several cases, has been shown to achieve exponentially lower label-complexity (number of queried labels) than supervised learning (Cohn et al., 1994). Domain Adaptation (Daumé & Marcu, 2006), although motivated somewhat differently, attempts to address a seemingly similar problem: lack of labeled data in some target domain. Domain Adaptation deals with this problem using labeled data from a different (but related) *source* domain.

In this paper, we consider the *supervised* domain adaptation setting (Finkel & Manning, 2009; Daumé III, 2007) having a large amount of labeled data from a source domain, a large amount of unlabeled data from a target domain, and *additionally* a small budget for acquiring labels in the target domain. We show how, apart from leveraging information in the usual domain adaptation sense, the information from the source domain can be leveraged to intelligently query labels in the target domain.We achieve this by first training the best possible classifier *without* using target domain labeled data [1] and then using the learned classifier to leverage the inter-domain information when we are additionally provided some fixed budget for acquiring extra *labeled* target data (i.e., the active learning setting (Settles, 2009)).

There are several ways in which our "best classifier" can be utilized. Our first approach uses this classifier as the initializer while doing (online) active learning in the target domain (Section 3). Then we present a variant augmenting the first approach using a domain-separator hypothesis which leads to *additionally* ruling out querying the labels of those target examples that appear "similar" to the source domain (Section 4).

Figure 1 shows our basic setup which uses a source (or unsupervised domain-adapted source) classifier $\mathbf{v}_0$ as an initializer for doing active learning in the target domain having some small, fixed budget for querying labels. Our framework consists of 2 phases: 1) Learning the best possible classi-

[1] For instance, either by simply training a supervised classifier on the labeled source data, or by using *unsupervised* domain adaptation techniques (Blitzer et al., 2006; Sugiyama et al., 2007) that use labeled data from the source domain, and additionally *unlabeled* data from the source and target domains.

*Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pages 27–32,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

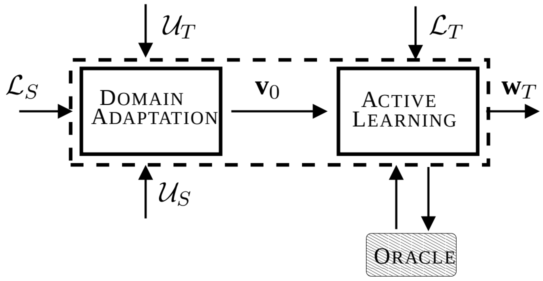

Figure 1: Block diagram of our basic approach. Stage-1 can use any black-box unsupervised domain adaptation approach (e.g., (Blitzer et al., 2006; Sugiyama et al., 2007))

fier $\mathbf{v}_0$ using source labeled ($\mathcal{L}_S$) and unlabeled data ($\mathcal{U}_S$), and target unlabeled ($\mathcal{U}_T$) data, and 2) Querying labels for target domain examples by leveraging information from the classifier learned in phase-1.

## 2 Online Active Learning

The active learning phase of our algorithm is based on (Cesa-Bianchi et al., 2006), henceforth referred to as CBGZ. In this section, we briefly describe this approach for the sake of completeness.

Their algorithm (Algorithm 1) starts with a zero initialized weight vector $\mathbf{w}_T^0$ and proceeds in rounds by querying the label of an example $x_i$ with probability $\frac{b}{b+|r_i|}$, where $|r_i|$ is the *confidence* (in terms of margin) of the current weight vector on $x_i$. $b$ is a parameter specifying how aggressively the labels are queried. A large value of $b$ implies that a large number of labels will be queried (conservative sampling) whereas a small value would lead to a small number of examples being queried. For each label queried, the algorithm updates the current weight vector if the label was predicted incorrectly. It is easy to see that the total number of labels queried by this algorithm is $\sum_{i=1}^{T} \mathbb{E}[\frac{b}{b+|r_i|}]$.

## 3 Active Online Domain Adaptation

In our supervised domain adaptation setting, we are given a small budget for acquiring labels in a target domain, which makes it imperative to use active learning in the target domain. However, our goal is to *additionally* also leverage inter-domain relatedness by exploiting whatever information we might already have from the source domain. To accomplish this, we take the online active learning approach of (Cesa-Bianchi et al., 2006) described in Section 2 and adapt it such that the algorithm uses the best possible classifier learned (*without* target labeled data; see Figure 1) as the initializer hypothesis in the target domain, and thereafter updates this hypothesis in an online fashion using actively acquired labels as is done in (Cesa-Bianchi et al., 2006). This amounts to using $\mathbf{w}_T^0 = \mathbf{v}_0$ in Algorithm 1. We refer to this algorithm as Active Online Domain Adaptation (AODA). It can be shown that the modified algorithm (AODA) yields smaller mistake bound and smaller label complexity than the CBGZ algorithm. We skip the proofs here and reserve the presentation for a longer version. It is however possible to provide an intuitive argument for the smaller label complexity: Since AODA is initialized with a non-zero (*but not randomly chosen*) hypothesis $\mathbf{v}_0$ learned using data from a related source domain, the sequence of hypotheses AODA produces are expected to have higher confidences margins $|r_i'|$ as compared that of CBGZ which is based on a *zero initialized hypothesis*. Therefore, at each round, the sampling probability of AODA given by $\frac{b}{b+|r_i'|}$ will also be smaller, leading to a smaller number of queried labels since it is nothing but $\sum_{i=1}^{T} \mathbb{E}[\frac{b}{b+|r_i'|}]$.

**Algorithm 1** CBGZ

```
**Input:** $b > 0$; $T$: number of rounds
**Initialization:** $\mathbf{w}_T^0 = 0$; $k = 1$;
**for** $i = 1$ to $T$ **do**
    $\hat{x}_i = x_i/||x_i||$, set $r_i = \mathbf{w}_T^{i-1}\hat{x}_i$;
    predict $\hat{y}_i = SIGN(r_i)$;
    sample $Z_i \sim Bernoulli(\frac{b}{b+|r_i|})$;
    **if** $Z_i = 1$ **then**
        query label $y_i \in \{+1, -1\}$
        **if** $\hat{y}_i \neq y_i$ **then**
            update: $\mathbf{w}_T^k = \mathbf{w}_T^{k-1} + y_i\hat{x}_i$; $k \leftarrow k + 1$;
        **end if**
    **end if**
**end for**
```

## 4 Using Domain Separator Hypothesis

The relatedness of source and target domains can be additionally leveraged to *further* improve the algorithm described in Section 3. Since the source and target domains are assumed to be related, one can use this fact to upfront rule out acquiring the labels of some target domain examples that “appear” to be similar to the source domain examples. As an illustration, Fig. 2 shows a typical distribution separator hypothesis (Blitzer et al., 2007a) which separates the *source* and *target* examples. If the source and target domains are reasonably different, then the separator hypothesis can perfectly distinguish between the examples drawn from these two domains. On the other hand, if the domains are similar, one would expected that there will be some overlap and therefore some of the target domain examples will lie on the source side (cf., Fig. 2). Acquiring labels for such examples is not really needed since the initializing hypothesis $\mathbf{v}_0$ (cf., Fig 1) of AODA would already have taken into account such examples. Therefore, such target examples can be outrightly ignored from being queried for labels. Our second algorithm (Algorithm 2) is similar to Algorithm 1, but also makes use of the distribution separator hypothesis (which can be learned using source and target *unlabeled* examples) as a preprocessing step before doing active learning on each incoming target example. We denote this algorithm by DS-AODA (for Domain-Separator based AODA). Since some of the target examples are upfront ruled out from being queried, this approach resulted even smaller number of queried labels (Section 5.4).

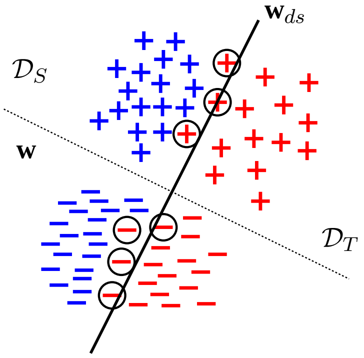

Figure 2: An illustrative diagram showing distribution separator hypothesis $\mathbf{w}_{ds}$ separating source data from target data. $\mathbf{w}$ is the actual target hypothesis

**Algorithm 2** DS-AODA

```
Input: $b > 0$; $\mathbf{w}_{ds}$: distribution separator hypothesis; $\mathbf{v}_0$ : initializing hypothesis ; $T$: number of rounds
Initialization: $\mathbf{w}_T^0 = \mathbf{v}_0$; $k = 1$;
for $i = 1$ to $T$ do
    $\hat{x}_i = x_i/||x_i||$,
    if $\hat{x}_i$ does not lie on the source side of $\mathbf{w}'_{ds}$ then
        set $r_i = \mathbf{w}_T^{i-1}\hat{x}_i$;
        predict $\hat{y}_i = SIGN(r_i)$;
        sample $Z_i \sim Bernoulli(\frac{b}{b+|r_i|})$;
        if $Z_i = 1$ then
            query label $y_i \in \{+1, -1\}$
            if $\hat{y}_i \neq y_i$ then
                update: $\mathbf{w}_T^k = \mathbf{w}_T^{k-1} + y_i\hat{x}_i$; $k \leftarrow k+1$;
            end if
        end if
    end if
end for
```

## 5 Experiments

In this section, we demonstrate the empirical performance of our algorithms and compare them with a number of baselines. Table 1 summarizes the methods used with a brief description of each. Among the first three (ID, sDA, FEDA), FEDA (Daumé III, 2007) is a state-of-the-art *supervised* domain adaptation method but assumes *passively* acquired labels. The last four, RIAL, ZIAL, SIAL and AODA methods in Table 1 acquire labels in an active fashion. As the description denotes, RIAL and ZIAL start active learning in *target* with a randomly initialized and zero initialized base hypothesis, respectively. It is also important to distinguish between SIAL and AODA here: SIAL uses an unmodified classifier learned only from *source* labeled data as the initializer, whereas AODA uses an *unsupervised* domain-adaptation technique (i.e., without using labeled target data) to learn the initializer. In our experiments, we use the instance reweighting approach (Sugiyama et al., 2007) to perform the unsupervised domain adaptation step. However, we note that this step can also be performed using any other unsupervised domain adaptation technique such as Structural Correspondence Learning (SCL) (Blitzer et al., 2006). We compare all the approaches based on classification accuracies achieved for a given budget of labeled target examples (Section-5.2), and number of labels requested for a fixed pool of unlabeled target examples and corresponding accuracies

| **Method** | **Summary** | **Active ?** |
|---|---|---|
| ID | In-domain ($\mathcal{D}_\mathcal{T}$) data | No |
| SDA | UDA followed by *passively* chosen labeled target data | No |
| FEDA | Frustratingly Easy Domain Adaptation Daumé III (2007) | No |
| ZIAL | Zero initialized active learning Cesa-Bianchi et al. (2006) | Yes |
| RIAL | Randomly initialized active learning with fixed label budget | Yes |
| SIAL | Source hypothesis initialized active learning | Yes |
| AODA | UDA based source hypothesis initialized active learning | Yes |

Table 1: Description of the methods compared

(Section-5.3). We use the vanilla Perceptron as the base classifier of each of the algorithms and each experiment has been averaged over 20 runs corresponding to random data order permutations.

### 5.1 Datasets

We report our empirical results for the task of sentiment classification using data provided by (Blitzer et al., 2007b) which consists of user reviews of eight product types (apparel, books, DVD, electronics, kitchen, music, video, and other) from Amazon.com. We also apply PCA to reduce the datadimensionality to 50. The sentiment classification task for this dataset is binary classification which corresponds to classifying a review as positive or negative. The sentiment dataset consists of several domain pairs with varying $\mathcal{A}$-distance (which measures the domain separation), akin to the sense described in (Ben-David et al., 2006). Table 2 presents the domain pairs used in our experiments and their corresponding domain divergences in terms of the $\mathcal{A}$-distance (Ben-David et al., 2006).

To compute the $\mathcal{A}$-distance from finite samples of source and target domain, we use a surrogate to the true $\mathcal{A}$-distance (the *proxy* $\mathcal{A}$-distance) in a manner similar to (Ben-David et al., 2006): First, we train a linear classifier to separate the *source* domain from the *target* domain using only unlabeled examples from both. The average per-instance hinge-loss of this classifier subtracted from 1 serves as our estimate of the *proxy* $\mathcal{A}$-distance. A score of 1 means perfectly separable distributions whereas a score of 0 means that the two distributions are essentially the same. As a general rule, a high score means that the two domains are reasonably far apart.

| **Source** | **Target** | **$\mathcal{A}$-distance** |
|---|---|---|
| Dvd (D) | Book (B) | 0.7616 |
| Dvd (D) | Music (M) | 0.7314 |
| Books (B) | Apparel (A) | 0.5970 |
| Dvd (D) | Apparel (A) | 0.5778 |
| Electronics (E) | Apparel (A) | 0.1717 |
| Kitchen (K) | Apparel (A) | 0.0459 |

Table 2: Proxy $\mathcal{A}$-distances between some domain pairs

### 5.2 Classification Accuracies

In our first experiment, we compare our first approach of Section 3 (AODA, and also SIAL which naïvely uses the *unadapted* source hypothesis) against other baselines on two domain pairs from the sentiments dataset: DVD→BOOKS (large $\mathcal{A}$ distance) and KITCHEN→APPAREL (small $\mathcal{A}$ distance) with varying target budget (1000 to 5000). The results are shown in Table 3 and Table 4. As the results indicate, on both datasets, our approaches (SIAL, AODA) perform consistently better than the baseline approaches (Table 1) which also include one of the state-of-the-art supervised domain adaptation algorithms (Daumé III, 2007). On the other hand, we observe that the zero-initialized and randomly initialized approaches do not perform as well. In particular, the latter case suggests that it's important to have a sensible initialization.

### 5.3 Label Complexity Results

Next, we compare the various algorithms on the basis of the number of labels acquired (and corresponding accuracies) when given the complete pool of unlabeled examples from the target domain. Table 5 shows that our approaches result in much smaller label complexities as compared to other ac-

| **Met-<br>hod** | **Target Budget** | | | | |
|---|---|---|---|---|---|
| | **1000** | **2000** | **3000** | **4000** | **5000** |
| | Acc (Std) | Acc (Std) | Acc (Std) | Acc (Std) | Acc (Std) |
| ID | 65.94 (±3.40) | 66.66 (±3.01) | 67.00 (±2.40) | 65.72 (±3.98) | 66.25 (±3.18) |
| sDA | 66.17 (±2.57) | 66.45 (±2.88) | 65.31 (±3.13) | 66.33 (±3.51) | 66.22 (±3.05) |
| RIAL | 51.79 (±4.36) | 53.12 (±4.65) | 55.01 (±4.20) | 57.56 (±4.18) | 58.57 (±2.44) |
| ZIAL | 66.24 (±3.16) | 66.72 (±3.30) | 63.97 (±4.82) | 66.28 (±3.61) | 66.36 (±2.82) |
| **SIAL** | **68.22 (±2.17)** | **69.65 (±1.20)** | **69.95 (±1.55)** | 70.54 (±1.42) | **70.97 (±0.97)** |
| **AODA** | 67.64 (±2.35) | 68.89 (±1.37) | 69.49 (±1.63) | **70.55 (1.15)** | 70.65 (±0.94) |
| FEDA | 67.31 (±3.36) | 68.47 (±3.15) | 68.37 (±2.72) | 66.95 (3.11) | 67.13 (±3.16) |
| **Acc:** Accuracy \| **Std:** Standard Deviation | | | | | |

Table 3: Classification accuracies for DVD→BOOKS, for fixed target budget.

| **Met-<br>hod** | **Target Budget** | | | | |
|---|---|---|---|---|---|
| | **1000** | **2000** | **3000** | **4000** | **5000** |
| | Acc (Std) | Acc (Std) | Acc (Std) | Acc (Std) | Acc (Std) |
| ID | 69.64 (±3.14) | 69.61 (±3.17) | 69.36 (±3.14) | 69.77 (±3.58) | 70.77 (±3.05) |
| sDA | 69.70 (±2.57) | 70.48 (±3.42) | 70.29 (±2.56) | 70.86 (±3.16) | 70.71 (±3.65) |
| RIAL | 52.13 (±5.44) | 56.83 (±5.36) | 58.09 (±4.09) | 59.82 (±4.16) | 62.03 (±2.52) |
| ZIAL | 70.09 (±3.74) | 69.96 (±3.27) | 68.6 (±3.94) | 70.06 (±2.84) | 69.75 (±3.26) |
| **SIAL** | 73.82 (±1.47) | **74.45 (±1.27)** | 75.11 (±0.98) | 75.35 (±1.30) | 75.58 (±0.85) |
| **AODA** | **73.93 (±1.84)** | 74.18 (±1.85) | **75.13 (±1.18)** | **75.88 (±1.32)** | **76.02 (±0.97)** |
| FEDA | 70.05 (±2.47) | 69.34 (±3.50) | 71.22 (±3.00) | 71.67 (±2.59) | 70.80 (±3.89) |
| **Acc:** Accuracy \| **Std:** Standard Deviation | | | | | |

Table 4: Classification accuracies for KITCHEN→APPAREL, for fixed target budget.

tive learning based baselines and still gives better classification accuracies. We also note that although RIAL initializes with a non-zero hypothesis and queries almost similar number of labels as our algorithms, it actually performs worse than even ZIAL in terms of classification accuracies, which implies the significant of a sensible initializing hypothesis.

| **Met-<br>hod** | **DVD→BOOK** | | **KITCHEN→APPAREL** | |
|---|---|---|---|---|
| | Acc (Std) | Labels | Acc (Std) | Labels |
| RIAL | 62.74 (±3.00) | 7618 | 62.15 (±4.51) | 4871 |
| ZIAL | 65.65 (±2.82) | 10459 | 70.19 (±2.64) | 6968 |
| **SIAL** | 72.11 (±1.20) | 7517 | 75.62 (±1.14) | **4709** |
| **AODA** | 72.00 (±1.31) | **7452** | 75.62 (±0.82) | 4752 |
| **Acc:** Accuracy \| **Std:** Standard Deviation | | | | |

Table 5: Accuracy and label complexity of DVD→BOOKS and KITCHEN→APPAREL *with full target training data treated as the unlabeled pool.*

### 5.4 DS-AODA Results

Finally, we evaluate our distribution separator hypothesis based approach (DS-AODA) discussed in Section 4. As our experimental results (on four domain pairs, Fig. 3) indicate, this approach leads to considerably smaller number of labels acquired than our first approach AODA which does not use the information about domain separation, without any perceptible loss in classification accuracies. Similar improvements in label complexity (although not reported here) were observed when we grafted the distribution separator hypothesis around SIAL (the unaltered source initialized hypothesis).

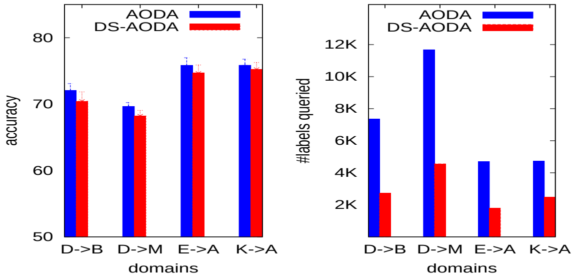

Figure 3: Test accuracy and label complexity of D→B, D→M, E→A and K→A.

### 5.5 SIAL vs AODA

Some of the results might indicate from naïvely initializing using even the unadapted source trained classifier (SIAL) tends to be as good as initializing with a classifer trained using unsupervised domain adaptation (AODA). However, it is mainly due to the particular unsupervised domain adaptation technique (naïve instance weighting) we have used here for the first stage. In some cases, the weights estimated using instance weighting may not be accurate and the bias in importance weight estimation is potentially the reason behind AODA not doing better than SIAL in such cases. As mentioned earlier, however, any other unsupervised domain adaptation technique can be used here and, in general, AODA is expected to perform better than SIAL.

## 6 Related Work

Active learning in a domain adaptation setting has received little attention so far. One interesting setting was proposed in (Chan & Ng, 2007) where they apply active learning for word sense disambiguation in a domain adaptation setting. Their active learning setting is pool-based whereas ours is a streaming (online) setting. Furthermore, our second algorithm also uses the domain separator hypothesis to rule out querying the labels of target examples similar to the source. A combination of transfer learning with active learning has been presented in (Shi et al., 2008). One drawback of their approach is the requirement of an initial pool of labeled target domain data used to train an in-domain classifier. Without this in-domain classifier, no transfer learning is possible in their setting.

## 7 Discussion

There are several interesting variants of our approach that can worth investigating. For instance, one can use a hybrid oracle setting where the source classifier $\mathbf{v}_0$ could be used as an oracle that provides labels for free, whenever it is reasonably highly confident about its prediction (maybe in terms of its relative confidence as compared to the actual classifier being learned; it would also be interesting to set, and possibly adapt, this confidence measure as the active learning progresses). Besides, in the distribution separator hypothesis based approach of Section 4, we empirically observed significant reductions in label-complexity, and it is supported by intuitive arguments. However, it would be interesting to be able to precisely quantify the amount by which the label-complexity is expected to reduce.

## References

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of Representations for Domain Adaptation. In *NIPS*, 2006.

Blitzer, J., Mcdonald, R., and Pereira, F. Domain Adaptation with Structural Correspondence Learning. In *EMNLP*, 2006.

Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Wortman, J. Learning Bounds for Domain Adaptation. In *NIPS*, 2007a.

Blitzer, J., Dredze, M., and Pereira, F. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *ACL*, 2007b.

Cesa-Bianchi, Nicolò, Gentile, Claudio, and Zaniboni, Luca. Worst-Case Analysis of Selective Sampling for Linear Classification. *JMLR*, 7, 2006.

Chan, Y. S. and Ng, H. T. Domain adaptation with active learning for word sense disambiguation. In *ACL*, 2007.

Cohn, David, Atlas, Les, and Ladner, Richard. Improving Generalization with Active Learning. *Machine Learning*, 15(2), 1994.

Daumé, III, Hal and Marcu, Daniel. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26(1), 2006.

Daumé III, H. Frustratingly Easy Domain Adaptation. In *ACL*, 2007.

Finkel, Jenny Rose and Manning, Christopher D. Hierarchical Bayesian domain adaptation. In *NAACL*, pp. 602–610, Morristown, NJ, USA, 2009.

Settles, B. Active Learning Literature Survey. In *Computer Sciences Technical Report 1648*, University of Wisconsin-Madison, 2009.

Shi, Xiaoxiao, Fan, Wei, and Ren, Jiangtao. Actively Transfer Domain Knowledge. In *ECML/PKDD (2)*, 2008.

Sugiyama, M., Nakajima, S., Kashima, H., von Bünau, P., and Kawanabe, M. Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation. In *NIPS*, 2007.