Online Active Learning for Cost Sensitive Domain Adaptation

Min Xiao and **Yuhong Guo**
Department of Computer and Information Sciences
Temple University
Philadelphia, PA 19122, USA
{minxiao,yuhong}@temple.edu

Abstract

Active learning and domain adaptation are both important tools for reducing labeling effort to learn a good supervised model in a target domain. In this paper, we investigate the problem of online active learning within a new active domain adaptation setting: there are insufficient labeled data in both source and target domains, but it is cheaper to query labels in the source domain than in the target domain. Given a total budget, we develop two cost-sensitive online active learning methods, a multi-view uncertainty-based method and a multi-view disagreement-based method, to query the most informative instances from the two domains, aiming to learn a good prediction model in the target domain. Empirical studies on the tasks of cross-domain sentiment classification of Amazon product reviews demonstrate the efficacy of the proposed methods on reducing labeling cost.

1 Introduction

In many application domains, it is difficult or expensive to obtain labeled data to train supervised models. It is critical to develop effective learning methods to reduce labeling effort or cost. Active learning and domain adaptation are both important tools for reducing labeling cost on learning good supervised prediction models. Active learning reduces the cost of labeling by selecting the most informative instances to label, whereas domain adaptation obtains auxiliary label information by exploiting labeled data in related domains. Combining the efforts from both areas to further reduce the labeling cost is an important research direction to explore.

In this paper, we consider online active learning with domain adaptations. Online learning has been widely studied (Borodin and El-Yaniv, 1998) due to its advantages of low memory requirement and fast computation speed. Dredze and Crammer (2008) applied online learning on domain adaptation and proposed to combine multiple similar source domains to perform online learning for the target domain, which provides a new opportunity for conducting active learning with domain adaptation. Online active learning with domain adaptation, to our knowledge, has just gained attention recently and has been addressed in (Rai et al., 2010; Saha et al., 2011). The active online domain adaptation methods developed in (Rai et al., 2010; Saha et al., 2011) leverage information from the source domain by domain adaptation to intelligently query labels for instances only in the target domain in an online fashion with a given budget. They assumed a large amount of labeled data is readily available in the source domain.

In this work, we however tackle online active learning with domain adaptation in a different setting, where source domains with a large amount of free labeled data are not available. Instead we assume there are very few labeled instances in both the source and target domains and labels in both domains can be acquired with a cost. Moreover, we assume the annotation cost for acquiring labels in the source domain is much lower than the annotation cost in the target domain. This is a practical setting in many domain adaptation scenarios. For example, one aims to learn a good review classification model for high-end computers. It may be expensive to acquire labels for such product reviews. However, but it might be relatively much cheaper (but not free) to acquire labels for reviews on movies or restaurants. In such an active learning scenario, will a source domain with lower annotation cost still be helpful for reducing the labeling cost required to learn a good prediction model in the target domain? Our research result in this paper will answer this ques-

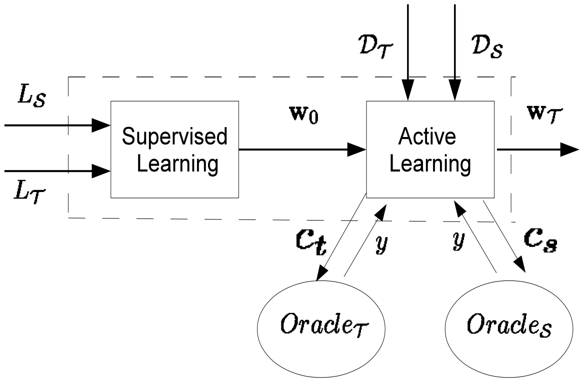


Figure 1: The framework of online active learning with domain adaptation.

tion. Specifically, we address this online active domain adaptation problem by extending the online active learning framework in (Cesa-Bianchi et al., 2006) to consider active label acquirement in both domains. We first initialize the prediction model based on the initial labeled data in both the source and target domains ($L_\mathcal{S}$ and $L_\mathcal{T}$). Then in each round of the online learning, we receive one unlabeled instance from each domain ($\mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{T}$), on which we need to decide whether to query labels. Whenever a label is acquired, we update the prediction model using the newly labeled instance if necessary. The framework of this online active learning setting is demonstrated in Figure 1. We exploit multi-view learning principles to measure the informativeness of instances and propose two cost-sensitive online active learning methods, a multi-view uncertainty-based method and a multi-view disagreement-based method, to acquire labels for the most informative instances. Our empirical studies on the tasks of cross-domain sentiment classification of Amazon product reviews show the proposed methods can effectively acquire the most informative labels given a budget, comparing to alternative methods.

2 Related Work

The proposed work in this paper involves research developments in multiple areas, including online active learning, active domain adaptation and multi-view active learning. In this section, we will cover the most related work in the literature.

Online active learning has been widely studied in the literature, including the perceptron-type methods in (Cesa-Bianchi et al., 2006; Monteleoni and Kääriäinen, 2007; Dasgupta et al., 2009). Cesa-Bianchi et al. (2006) proposed a selective sampling perceptron-like method (CBGZ), which serves as a general framework of online active learning. Monteleoni and Kääriäinen (2007) empirically studied online active learning algorithms, including the CBGZ, for optical character recognition applications. Dasgupta et al. (2009) analyzed the label complexity of the perceptron algorithm and presented a combination method of a modification of the perceptron update with an adaptive filtering rule. Our proposed online active learning methods are placed on an extended framework of (Cesa-Bianchi et al., 2006), by incorporating domain adaptation and multi-view learning techniques in an effective way.

Active domain adaptation has been studied in (Chan and Ng, 2007; Rai et al., 2010; Saha et al., 2011; Li et al., 2012). Chan and Ng (2007) presented an early study on active domain adaptation and empirically demonstrated that active learning can be successfully applied on out-of-domain word sense disambiguation systems. Li et al. (2012) proposed to first induce a shared subspace across domains and then actively label instances augmented with the induced latent features. Online active domain adaptation, however, has only been recently studied in (Rai et al., 2010; Saha et al., 2011). Nevertheless, the active online domain adaptation method (AODA) and its variant method, domain-separator based AODA (DS-AODA), proposed in these works assume a large amount of labeled data in the source domain and conduct online active learning only in the target domain, which is different from our problem setting in this paper.

Multi-view learning techniques have recently been employed in domain adaptation (Tur, 2009; Blitzer et al., 2011; Chen et al., 2011). In particular, instead of using data with conditional independent views assumed in standard multi-view learning (Blum and Mitchell, 1998), Blitzer et al. (2011) and Chen et al. (2011) randomly split original features into two disjoint subsets to produce two views, and demonstrate the usefulness of multi-view learning with synthetic two views. On the other hand, *multi-view active learning* has been studied in (Muslea et al., 2000, 2002; Wang and Zhou, 2008, 2010). These works all suggest to query labels for *contention points* (instances on which different views predict different labels). Our proposed methods will exploit this multi-view principle and apply it in our multi-view online active domain adaptation setting.

In addition, our proposed work is also related to *cost-sensitive active learning*. But different from the traditional cost-sensitive active learning, which assumes multiple oracles with different costs exist for the same set of instances (Donmez and Carbonell, 2008; Arora et al., 2009), we assume two oracles, one for the source domain and one for the target domain. Overall, the problem we study in this paper is novel, practical and important. Our research will demonstrate a combination of advances in multiple research areas.

3 Multi-View Online Active Learning with Domain Adaptation

Our online active learning is an extension of the online active perceptron learning framework of (Cesa-Bianchi et al., 2006; Rai et al., 2010) in the cost-sensitive online active domain adaption setting. We will present two multi-view online active methods in this section under the framework shown in Figure 1.

Assume we have a target domain ($\mathcal{D}_\mathcal{T}$) and a related source domain ($\mathcal{D}_\mathcal{S}$) with a few labeled instances, $L_\mathcal{T}$ and $L_\mathcal{S}$, in each of them respectively. The instances in the two domains are drawn from the same input space but with two different distributions specified by each domain. An initial prediction model ($\mathbf{w_0}$) can then be trained with the current labeled data from both domains. Many domain adaptation techniques (Sugiyama, 2007; Blitzer et al., 2011) can be used for training here. However, for simplicity of demonstrating the effectiveness of online active learning strategies, we use vanilla Perceptron to train the initial prediction model on all labeled instances, as the perceptron algorithm is widely used in various works (Saha et al., 2011) and can be combined seamlessly with the online perceptron updates. It can be viewed as a simple supervised domain adaptation training.

The very few initial labeled instances are far from being sufficient to train a good prediction model in the target domain. Additional labeled data needs to be acquired to reach a reasonable prediction model. However it takes time, money, and effort to acquire labels in all problem domains. For simplicity of demonstration, we use money to measure the cost and effort of labeling instances in each domain. Assume the cost of labeling one instance in the source domain is c_s and the cost of labeling one instance in the target domain is c_t, where $c_t > c_s$. Note the condition $c_t > c_s$ is one criterion to be guaranteed when selecting useful source domains. It does not make sense to select source domains with more expensive labeling cost. Given a budget $\mathbf{B}$, we need to make wise decisions about which instances to query in the online learning setting. We aim to learn the best prediction model in the target domain with the labels purchased under the given budget.

Then online active learning will be conducted in a sequence of rounds. In each round r, we will receive two randomly sampled unlabeled instances in parallel, $\mathbf{x}_{s,r}$ and $\mathbf{x}_{t,r}$, one from each domain, $\mathbf{x}_{s,r} \in \mathcal{D}_\mathcal{S}$ and $\mathbf{x}_{t,r} \in \mathcal{D}_\mathcal{T}$. Active learning strategies will be used to judge the informativeness of the two instances in a cost-sensitive manner and decide whether to query labels for any one of them to improve the prediction model in the target domain. After new labels being acquired, we use the newly labeled instances to make online perceptron updates if the true labels are different from the predicted labels.

In this work, we focus on binary prediction problems where the labels have binary values, $y \in \{+1, -1\}$. We adopt the online perceptron-style learning model of (Cesa-Bianchi et al., 2006) for the online updates of the supervised perceptron model. Moreover, we extend principles of multi-view active learning into our online active learning framework. As we introduced before, synthetic multi-views produced by splitting the original feature space into disjoint subsets have been demonstrated effective in a few previous work (Blitzer et al., 2011; Chen et al., 2011). We adopt this idea to generate two views of the instances in both domains by randomly splitting the common feature space into two disjoint feature subsets, such that $\mathbf{x}_{s,r} = \{\mathbf{x}_{s,r}^{(1)}, \mathbf{x}_{s,r}^{(2)}\}$ and $\mathbf{x}_{t,r} = \{\mathbf{x}_{t,r}^{(1)}, \mathbf{x}_{t,r}^{(2)}\}$. Thus the initial prediction model will include two predictors $(f^{(1)}, f^{(2)})$ with model parameters $(\mathbf{w}_0^{(1)}, \mathbf{w}_0^{(2)})$, each trained on one view of the labeled data using the perceptron algorithm. Correspondingly, the online updates will be made on the two predictors.

The *critical challenge* of this cost-sensitive online active learning problem nevertheless lies in how to select the most informative instances for labeling. Based on different measurements of instance informativeness, we propose two online active learning algorithms: a Multi-view

Uncertainty-based instance Selection (MUS) algorithm and a Multi-view Disagreement-based instance Selection (MDS) algorithm for cost-sensitive online active domain adaptation, which we will present below.

3.1 Multi-View Uncertainty-based Instance Selection Algorithm

We use the initial model ($f^{(1)}, f^{(2)}$), trained on the two views of the initial labeled data and represented by the model parameters ($\mathbf{w}_0^{(1)}, \mathbf{w}_0^{(2)}$), as the starting point of the online active learning.

In each round r of the online active learning, we receive two instances $\mathbf{x}_{s,r} = \{\mathbf{x}_{s,r}^{(1)}, \mathbf{x}_{s,r}^{(2)}\}$ and $\mathbf{x}_{t,r} = \{\mathbf{x}_{t,r}^{(1)}, \mathbf{x}_{t,r}^{(2)}\}$, one for each domain. For the received instances, we need to make two sequential decisions:

1. Between the instance ($\mathbf{x}_{s,r}$) from the source domain and the instance ($\mathbf{x}_{t,r}$) from the target domain, which one should we select for further consideration?

2. For the selected instance, do we really need to query its label?

We answer the first question based on the labeling cost ratio, c_t/c_s, from the two domains and define the following probability

$$P_c = e^{-\alpha(c_t/c_s - 1)} \tag{1}$$

where α is a domain preference weighting parameter. Then with a probability P_c we select the target instance $\mathbf{x}_{t,r}$ and with a probability $1 - P_c$ we select the source instance $\mathbf{x}_{s,r}$. Our intuition is that one should query the less expensive source domain more frequently. Thus more labeled instances can be collected within the fix budget. On the other hand, the more useful and relevant but expensive instances from the target domain should also be queried at a certain rate.

For the selected instance $\mathbf{x}_{*,r}$, we then use a multi-view uncertainty strategy to decide whether to query its label. We first calculate the prediction confidence and predicted labels of the selected instance based on the current predictors trained from each view

$$m_k = |\mathbf{w}^{(k)\top}\mathbf{x}_{*,r}^{(k)}|, \;\; \widehat{y}^{(k)} = sign(\mathbf{w}^{(k)\top}\mathbf{x}_{*,r}^{(k)}) \tag{2}$$

where $k = 1$ or 2, standing for each of the two views. If the two predictors disagree over the prediction label, i.e., $\widehat{y}^{(1)} \neq \widehat{y}^{(2)}$, the selected instance is a contention point and contains useful information for at least one predictor, according to the principle of multi-view active learning. We then decide to pay a cost (c_s or c_t) to query its label. Otherwise, we make the query decision based on the two predictors' uncertainty (i.e., the inverse of the prediction confidence m_k) over the selected instance. Specifically, we sample two numbers, one for each view, according to

$$z_k = \text{Bernoulli}(b/(b + m_k)) \tag{3}$$

where b is a prior hyperparameter, specifying the tendency of querying labels. In our experiments, we use $b = 0.1$. If either $z_1 = 1$ or $z_2 = 1$, which means that at least one view is uncertain about the selected instance, we will query for the label y. The prediction model will be updated using the new labeled instances when the true labels are different from the predicted ones; i.e.,

$$\mathbf{w}^{(k)} = \mathbf{w}^{(k)} + (y\mathbf{x}_{*,r}^{(k)})I[y \neq \widehat{y}^{(k)}] \tag{4}$$

for $k = 1, 2$, where $I[\cdot]$ is an indicator function. This multi-view uncertainty-based instance selection algorithm (MUS) is given in Algorithm 1.

Algorithm 1 MUS Algorithm

```
Input: $\mathbf{B}, P_c, c_s, c_t, b,$
       initial model $(\mathbf{w}_0^{(1)}, \mathbf{w}_0^{(2)})$
Output: prediction model $(\mathbf{w}^{(1)}, \mathbf{w}^{(2)})$
Initialize: $\mathbf{w}^{(1)} = \mathbf{w}_0^{(1)}$, $\mathbf{w}^{(2)} = \mathbf{w}_0^{(2)}$
for each round $r = 1, 2, \cdots$ do
  Receive two instances $\mathbf{x}_{s,r},\ \mathbf{x}_{t,r}$
  Sample $d \sim U(0, 1)$
  if $\mathbf{B} < c_t$ then $d = 1$ end if
  if $d > P_c$ then $\mathbf{x}_{*,r}$= $\mathbf{x}_{s,r},\ c = c_s$
   else $\mathbf{x}_{*,r}$= $\mathbf{x}_{t,r},\ c = c_t$
  end if
  Compute $m_1, m_2, \widehat{y}^{(1)}, \widehat{y}^{(2)}$ by Eq.(2)
  Compute $z_1, z_2$ by Eq.(3)
  if $z_1 = 1$ or $z_2 = 1$ or $\widehat{y}^{(1)} \neq \widehat{y}^{(2)}$ then
   Query label $y$ for $\mathbf{x}_{*,r},\ \mathbf{B} = \mathbf{B} - c$
   Update $(\mathbf{w}^{(1)}, \mathbf{w}^{(2)})$ by Eq (4)
  end if
  if $\mathbf{B} < c_s$ then break end if
end for
```

3.2 Multi-View Disagreement-based Instance Selection Algorithm

MUS is restrained to query at most one instance at each round of the online active learning. In this section, we present an alternative multi-view disagreement-based instance selection algorithm (MDS) within the same framework.

In each round r of the online active learning, given the two instances $\mathbf{x}_{s,r}$ and $\mathbf{x}_{t,r}$ we received, the MDS algorithm evaluates both instances for potential label acquisition using the multi-view information provided by the two per-view predictors. Let $\widehat{y}_s^{(1)}$ and $\widehat{y}_s^{(2)}$ denote the predicted labels of instance $\mathbf{x}_{s,r}$ produced by the two predictors according to Eq (2). Similarly let $\widehat{y}_t^{(1)}$ and $\widehat{y}_t^{(2)}$ denote the predicted labels of instance $\mathbf{x}_{t,r}$. Follow the principle suggested in the multi-view active learning work (Muslea et al., 2000, 2002; Wang and Zhou, 2008, 2010) that querying labels for *contention points* (instances on which different views predict different labels) can lead to superior information gain than querying uncertain points, we identify the non-redundant contention points from the two domains for label acquisition.

Specifically, there are three cases: (1) If only one of the instances is a contention point, we query its label with probability P_c (Eq (1)) when the instance is from the target domain, and query its label with probability $1 - P_c$ when the instance is from the source domain. (2) If both instances are contention points, i.e., $\widehat{y}_s^{(1)} \neq \widehat{y}_s^{(2)}$ and $\widehat{y}_t^{(1)} \neq \widehat{y}_t^{(2)}$, but the predicted labels for the two instances are the same, i.e., $\widehat{y}_s^{(k)} = \widehat{y}_t^{(k)}$ for $k = 1, 2$, it suggests the two instances contain similar information with respect to the prediction model and we only need to query one of them. We then select the instance in a cost-sensitive manner stated in the MUS algorithm by querying the target instance with a probability P_c and querying the source instance with a probability $1 - P_c$. (3) If both instances are contention points but with different predicted labels, it suggests the two instances contain complementary information with respect to the prediction model, and we thus query labels for both of them.

For any new labeled instance from the target domain or the source domain, we update the prediction model of each review using Equation (4) when the acquired true label is different from the predicted label. The overall MDS algorithm is given in Algorithm 2.

Algorithm 2 MDS Algorithm

```
Input: $\mathbf{B}, P_c, c_s, c_t, b$,
       initial model $(\mathbf{w}_0^{(1)}, \mathbf{w}_0^{(2)})$
Output: prediction model $(\mathbf{w}^{(1)}, \mathbf{w}^{(2)})$
Initialize: $\mathbf{w}^{(1)} = \mathbf{w}_0^{(1)}$, $\mathbf{w}^{(2)} = \mathbf{w}_0^{(2)}$
for each round $r = 1, 2, \cdots$ do
  Receive two instances $\mathbf{x}_{s,r},\ \mathbf{x}_{t,r}$
  Compute $\widehat{y}_s^{(1)}, \widehat{y}_s^{(2)}, \widehat{y}_t^{(1)}, \widehat{y}_t^{(2)}$ by Eq (2)
  Let $d_s = I[\widehat{y}_s^{(1)} = \widehat{y}_s^{(2)}]$, $d_t = I[\widehat{y}_t^{(1)} = \widehat{y}_t^{(2)}]$
  Let $q_s = 0, q_t = 0$
  if $\mathbf{B} < c_t$ then $d_t = 0$ end if
  Sample $d \sim U(0, 1)$
  if $d_s = 1$ and $d_t = 0$ then
    if $d > P_c$ then $q_s = 1$ end if
  else if $d_s = 0$ and $d_t = 1$ then
    if $d \leq P_c$ then $q_t = 1$ end if
  else if $d_s = 1$ and $d_t = 1$
    if $\widehat{y}_s^{(1)} = \widehat{y}_t^{(1)}$ then
      if $d > P_c$ then $q_s = 1$ else $q_t = 1$ end if
    else $q_s = 1, q_t = 1$
    end if
  end if
  if $q_s = 1$ then
    Query label $y_s$ for $\mathbf{x}_{s,r},\ \mathbf{B} = \mathbf{B} - c_s$
    Update $(\mathbf{w}^{(1)}, \mathbf{w}^{(2)})$ by Eq (4)
  end if
  if $\mathbf{B} < c_t$ then $q_t = 0$ end if
  if $q_t = 1$ then
    Query label $y_t$ for $\mathbf{x}_{t,r},\ \mathbf{B} = \mathbf{B} - c_t$
    Update $(\mathbf{w}^{(1)}, \mathbf{w}^{(2)})$ by Eq (4)
  end if
  if $\mathbf{B} < c_s$ then break end if
end for
```

3.3 Multi-View Prediction

After the training process, we use the two predictors to predict labels of the test instances from the target domain. Given a test instance $\mathbf{x}_t = (\mathbf{x}_t^{(1)}, \mathbf{x}_t^{(2)})$, we use the predictor that have larger prediction confidence to determine its label y^*. The prediction confidence of the kth view predictor on $\mathbf{x}_t$ is defined as the absolute prediction value $|\mathbf{w}^{(k)\top}\mathbf{x}_t^{(k)}|$. We then select the most confident predictor for this instance as

$$k^* = \arg\max_{k \in \{1,2\}} |\mathbf{w}^{(k)\top}\mathbf{x}_t^{(k)}| \qquad (5)$$

The predicted label is final computed as

$$y^* = sign(\mathbf{w}^{(k^*)\top}\mathbf{x}_t^{(k^*)}) \qquad (6)$$

With this multi-view prediction on the test data, the multi-view strengths can be exploited in the testing phase as well.

4 Experiments

In this section, we present the empirical evaluation of the proposed online active learning methods on the task of sentiment classification comparing to alternative baseline methods. We first describe the experimental setup, and then present the results and discussions.

4.1 Experimental Setup

Dataset For the sentiment classification task, we use the dataset provided in (Prettenhofer and Stein, 2010). The dataset contains reviews with four different language versions and in three domains, *Books (B)*, *DVD (D)* and *Music (M)*. Each domain contains 2000 positive reviews and 2000 negative reviews, with a term-frequency (TF) vector representation. We used the English version and constructed 6 source-target ordered domain pairs from the original 3 domains: *B2D, D2B, B2M, M2B, D2M, M2D*. For example, for the task of *B2D*, we use the Books reviews as the source domain and the DVD reviews as the target domain. For each pair of domains, we built a unigram vocabulary over the combined 4000 source reviews and 4000 target reviews. We further preprocessed the data by removing features that appear less than twice in either domain, replacing TF with TFIDF, and normalizing the attribute values into $[0, 1]$.

Approaches In the experiments, we mainly compared the proposed *MUS* and *MDS* algorithms with the following three baseline methods. (1) *MTS* (Multi-view Target instance Selection): It is a target-domain variant of the MUS algorithm, and selects the most uncertain instance received from the target domain to query according to the procedure introduced for MUS method. (2) *TCS* (Target Contention instance Selection): It is a target-domain variant of the MDS algorithm, and uses multi-view predictors to query contention instances received from the target domain. (3) *SUS* (Single-view Uncertainty instance Selection): It selects target vs source instances according to P_c (see Eq.(1)), and then uses uncertainty measure to make query decision. This is a single view variant of the MUS algorithm. In the experiments, we used $\alpha = 1$ for the P_c computation in Eq.(1).

4.2 Classification Accuracy

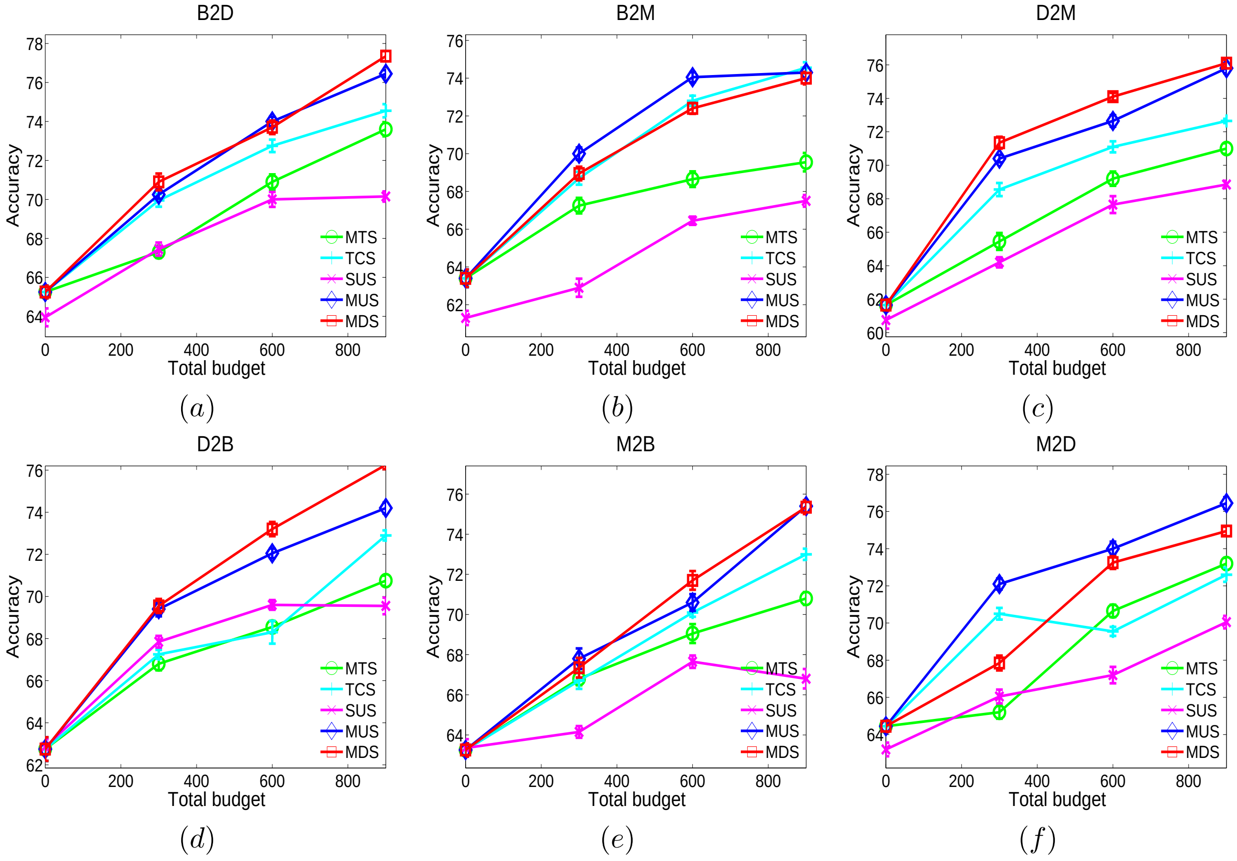
We first conducted experiments over the 6 domain adaptation tasks constructed from the sentiment classification data with a fixed cost ratio $c_t/c_s = 3$. We set $c_s = 1$ and $c_t = 3$. Given a budget $\mathbf{B} = 900$, we measure the classification performance of the prediction model learned by each online active learning method during the process of budget being used. We started with 50 labeled instances from the source domain and 10 labeled instances from the target domain. The classification performance is measured over 1000 test instances from the target domain. All other instances are used as inputs in the online process. We repeated the experiments 10 times using different random online instance input orders. The average results are reported in Figure 2.

The results indicate the proposed two algorithms, MUS and MDS, in general greatly outperform the other alternative methods. The SUS method, which is a single-view variant of MUS, presents very poor performance across all 6 tasks comparing to the other multi-view based methods, which demonstrates the efficacy of the multi-view instance selection mechanism. Among the multi-view based active learning methods, the MTS method and TCS method, which only query labels for more relevant but expensive instances from the target domain, demonstrated inferior performance, comparing to their cost-sensitive counterparts, MUS and MDS, respectively. This suggests that a cheaper source domain is in general helpful on reducing the labeling cost for learning a good prediction model in the target domain and our proposed active learning strategies are effective.

4.3 Domain Divergence

To further validate and understand our experimental results on the sentiment classification data, we evaluated the domain divergence over the three pairs of domains we used in the experiments above. Note, if the domain divergence is very small, it will be natural that a cheaper source domain should help on reducing the labeling cost in the target domain. If the domain divergence is very big, the space of exploring a cheaper source domain will be squeezed.

The divergence of two domains can be measured using the $\mathcal{A}$-distance (Ben-David et al., 2006). We adopted the method of (Rai et al., 2010) to proximate the $\mathcal{A}$-distance. We train a linear classifier over all 8000 instances, 4000 instances from each domain, to separate the two domains. The average per-instance hinge-loss for this separator subtracted from 1 was used as the estimate



(a) (b) (c)

(d) (e) (f)

Figure 2: Online active learning results over the 6 domain adaptation tasks for sentiment classification, with a total budget B=900 and a fixed cost ratio $c_t/c_s = 3$.

of the proxy $\mathcal{A}$-distance. A score of 1 means perfectly separable distributions and 0 means the two distributions from the two domains are identical. In general, a higher score means a larger divergence between the two domains.

Table 1: Proxy $\mathcal{A}$-distance over domain pairs.

Domains	**$\mathcal{A}$-distance**
Books vs. DVD	0.7221
Books vs. Music	0.8562
DVD vs. Music	0.7831

The proxy $\mathcal{A}$-distances over the 3 domain pairs from the sentiment classification dataset are reported in Table 1. It shows that all the 3 pairs of domains are reasonably far apart. This justified the effectiveness of the online active domain adaptation methods we developed and the results we reported above. It suggests the applicability of the proposed active learning scheme is not bound to the existence of highly similar source domains. Moreover, the $\mathcal{A}$-distance between *Books* and *Music* is the largest among the three pairs. Thus it is most challenging to exploit the source domain in the adaptation tasks, B2M and M2B. This explains the good performance of the target-domain method TCS on these two tasks. Nevertheless, the proposed MUS and MDS maintained consistent good performance even on these two tasks.

4.4 Robustness to Cost Ratio

We then studied the empirical behavior of the proposed online active domain adaptation algorithms with different cost ratio values c_t/c_s.

Given a fixed budget $B = 900$, we set $c_s = 1$ and run a few sets of experiments on the sentiment classification data by setting c_t as different values from $\{1, 2, 3, 4\}$, under the same experimental setting described above. In addition to the five comparison methods used before, we also added a baseline marker, *SCS*, which is a source-domain variant of the *MDS* algorithm and queries contention instances from only the source domain. The final classification performance of the prediction model learned with each approach is recorded

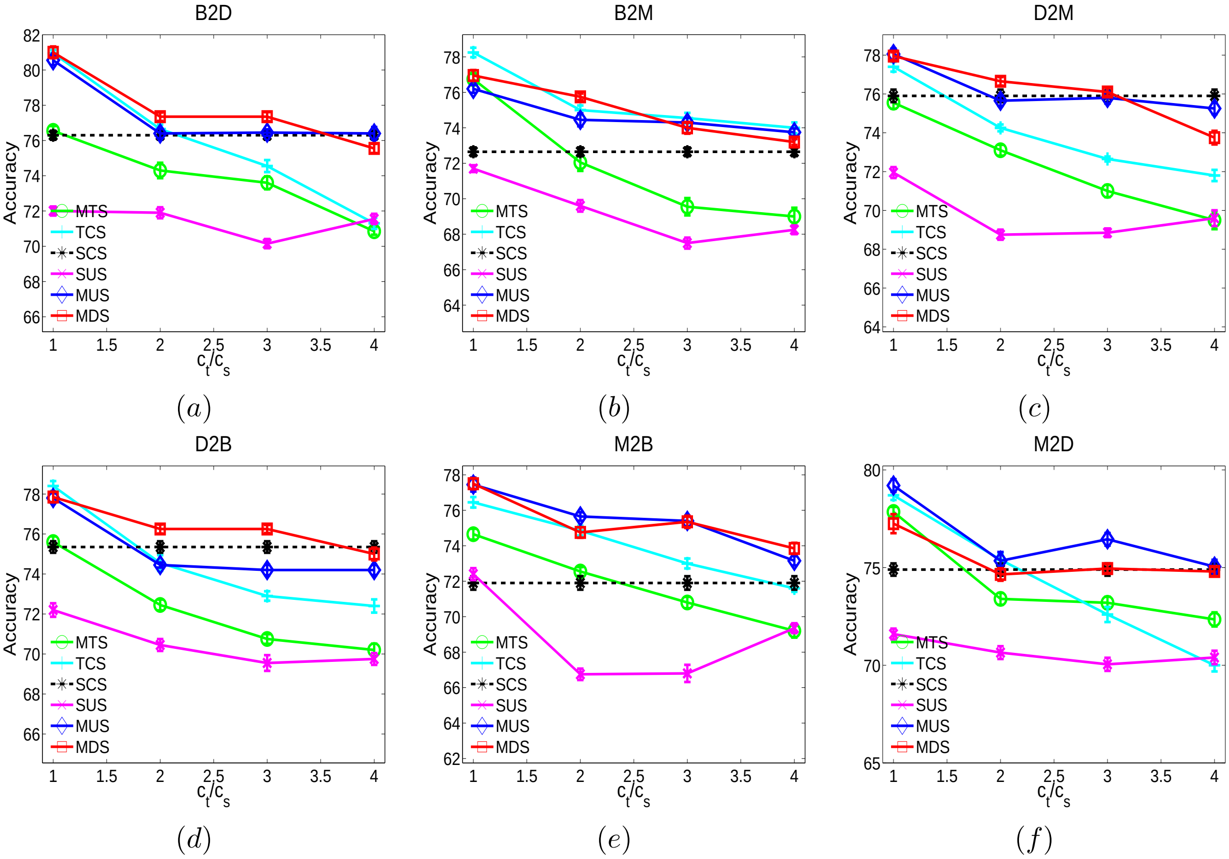


Figure 3: Online active learning results over the 6 domain adaptation tasks for sentiment classification, with different cost ratio values $c_t/c_s = \{1, 2, 3, 4\}$.

after the whole budget being used. The average results over 10 runs are reported in Figure 3.

We can see that: (1) With the increasing of the labeling cost in the target domain, the performance of all methods except *SCS* decreases since the same budget can purchase fewer labeled instances from the target domain. (2) The three cost-sensitive methods (SUS, MUS, and MDS), which consider the labeling cost when making query decisions, are less sensitive to the cost ratios than the MTS and TCS methods, whose performance degrades very quickly with the increasing of c_t/c_s. (3) It is reasonable that when c_t/c_s is very big, the SCS, which simply queries source instances, produces the best performance. But the proposed two cost-sensitive active learning methods, MUS and MDS, are quite robust to the cost ratios across a reasonable range of c_t/c_s values, and outperform both source-domain only and target-domain only methods. When $c_t = c_s$, the proposed cost-sensitive methods automatically favor target instances and thus achieve similar performance as TCS. When c_t becomes much larger than c_s, the proposed cost-sensitive methods automatically adjust to favor cheaper source instances and maintain their good performance.

5 Conclusion

In this paper, we investigated the online active domain adaptation problem in a novel but practical setting where we assume labels can be acquired with a lower cost in the source domain than in the target domain. We proposed two multi-view online active learning algorithms, MUS and MDS, to address the proposed problem. The proposed algorithms exploit multi-view active learning learning principles to measure the informativeness of instances and select instances in a cost-sensitive manner. Our empirical studies on the task of cross-domain sentiment classification demonstrate the efficacy of the proposed methods. This research shows that a cheaper source domain can help on reducing labeling cost for learning a good prediction model in the related target domain, with proper designed active learning algorithms.

Acknowledgments

This research was supported in part by NSF grant IIS-1065397.

References

S. Arora, E. Nyberg, and C. P. Rosé. Estimating annotation cost for active learning in a multi-annotator environment. In *Proc. of the NAACL-HLT 2009 Workshop on Active Learning for Natural Language Processing*, 2009.

S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

J. Blitzer, D. Foster, and S. Kakade. Domain adaptation with coupled subspaces. In *Proc. of the Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proc. of the Conference on Computational Learning Theory (COLT)*, 1998.

A. Borodin and R. El-Yaniv. *Online computation and competitive analysis*. Cambridge University Press, 1998.

N. Cesa-Bianchi, C. Gentile, and L. Zaniboni. Worst-case analysis of selective sampling for linear classification. *Journal of Machine Learning Research (JMLR)*, 7:1205–1230, 2006.

Y. Chan and H. Ng. Domain adaptation with active learning for word sense disambiguation. In *Proc. of the Annual Meeting of the Assoc. of Computational Linguistics (ACL)*, 2007.

M. Chen, K. Weinberger, and J. Blitzer. Co-training for domain adaptation. In *Advances in Neural Inform. Process. Systems (NIPS)*, 2011.

S. Dasgupta, A. T. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. *Journal of Machine Learning Research (JMLR)*, 10:281–299, 2009.

P. Donmez and J. G. Carbonell. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *Proc. of the ACM Conference on Information and knowledge management (CIKM)*, 2008.

M. Dredze and K. Crammer. Online methods for multi-domain learning and adaptation. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2008.

L. Li, X. Jin, S. Pan, and J. Sun. Multi-domain active learning for text classification. In *Proc. of the ACM SIGKDD Inter. Conference on Knowledge Discovery and Data Mining (KDD)*, 2012.

C. Monteleoni and M. Kääriäinen. Practical online active learning for classification. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Online Learning for Classification Workshop*, 2007.

I. Muslea, S. Minton, and C. Knoblock. Selective sampling with redundant views. In *Proc. of the National Conference on Artificial Intelligence (AAAI)*, 2000.

I. Muslea, S. Minton, and C. A. Knoblock. Active + semi-supervised learning = robust multi-view learning. In *Proc. of the International Conference on Machine Learning (ICML)*, 2002.

P. Prettenhofer and B. Stein. Cross-language text classification using structural correspondence learning. In *Proc. of the Annual Meeting for the Association of Computational Linguistics (ACL)*, 2010.

P. Rai, A. Saha, H. Daumé III, and S. Venkatasubramanian. Domain adaptation meets active learning. In *Proc. of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2010.

A. Saha, P. Rai, H. Daumé III, S. Venkatasubramanian, and S. DuVall. Active supervised domain adaptation. In *Proc. of the European Conference on Machine Learning (ECML)*, 2011.

M. Sugiyama. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

G. Tur. Co-adaptation: Adaptive co-training for semi-supervised learning. In *Proc. of the IEEE Inter. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.

W. Wang and Z. Zhou. On multi-view active learning and the combination with semi-supervised learning. In *Proc. of the international conference on Machine learning (ICML)*, 2008.

W. Wang and Z. Zhou. Multi-view active learning in the non-realizable case. In *Advances in Neural Inform. Process. Systems (NIPS)*, 2010.