Multi-Source Iterative Adaptation for Cross-Domain Classification

Himanshu S. Bhatt, Arun Rajkumar and Shourya Roy
Xerox Research Centre India, Bengaluru, INDIA
{Firstname.Lastname}@xerox.com

Abstract

Owing to the tremendous increase in the volume and variety of user generated content, *train–once–apply–forever* models are insufficient for supervised learning tasks. Thus, developing algorithms that adapt across domains by leveraging data from multiple domains is critical. However, existing adaptation algorithms often fail to identify the right sources to use for adaptation. In this work, we present a novel multi-source iterative domain adaptation algorithm (MSIDA) that leverages knowledge from selective sources to improve the performance in a target domain. The algorithm first chooses the best K sources from possibly numerous existing domains taking into account both similarity and complementarity properties of the domains. Then it learns target specific features in an iterative manner building on the common shared representations from the source domains. We give theoretical justifications for our source selection procedure and also give mistake bounds for the MSIDA algorithm. Experimental results justify the theory as MSIDA significantly outperforms existing cross-domain classification approaches on the real world and benchmark datasets.

1 Introduction

Internet and social media have led to the astronomical growth of user generated content in a plethora of domains. Analyzing this data is often time consuming or requires expensive manual annotations (say as positive, negative or neutral for the sentiment categorization task) to train machine learning models. With high velocity and variety of social media data, more and more domains and tasks are evolving. While traditional machine learning approaches would need the training and testing instances to be from the same distribution/domain (for e.g. train on reviews from books and test on reviews of books), transfer learning and domain adaptation methods allow domains, tasks and distributions used in training and testing to be different, but related. Consider a *multi-source cross-domain classification* scenario where we train on reviews from books, dvds and kitchen domains while test on reviews from electronics domain. Several domain adaptation techniques [Pan and Yang, 2010; Sun *et al.*, 2015] have been proposed in literature; however in our experience with real world data, we observe a few defining features of this problem which restrain the practicality of most of the existing algorithms.

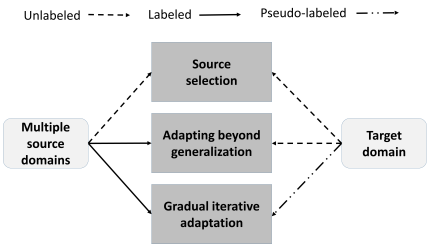

Figure 1: Lists the building blocks of the proposed algorithm.

- Multi-source domain adaptation has been explored in literature [Sun *et al.*, 2015]. However, when the number of sources is very large (especially in web based applications), it is important not only to adapt but also to adapt by *choosing a handful of good sources*. Choosing the best source domains to adapt from given a target domain is a critical issue that has got very little attention so far.
- The basic philosophy of domain adaptation is to *generalize and adapt*, i.e., generalize from existing domains and adapt to the target domain. Most domain adaptation algorithms hope to achieve generalization by learning some form of common shared representation. However, domain adaptation techniques prove inadequate to "adapt" to the target specific discriminating features.
- In practical domain adaptation scenarios, one of the big challenges is *negative transfer* which happens when knowledge transfer has a negative impact on target learning [Pan and Yang, 2010]. Traditional one-shot transfer learning approaches do not account for one of the fundamental questions that is *how much to transfer?* and thus, suffer due to aggressive transfer.

Our main contribution is to propose novel algorithms to address the above three fundamental challenges, illustrated as the building blocks of our approach in Figure 1. We first propose a novel source selection algorithm which finds the best K sources to adapt for a target domain given multi-

ple sources. Once the sources are selected, we propose the multi source domain adaptation algorithm (MSIDA) to learn target-specific discriminating features from unlabeled target domain starting with an initial common shared representation. MSIDA gradually and iteratively transfers knowledge using an ensemble of shared and target specific classifiers, instead of the more common one-shot transfers.

2 Related Work

Domain adaptation has been used in several applications including text classification [Chen *et al.*, 2009; Dai *et al.*, 2007], part of speech tagging [Ando and Zhang, 2005; Daumé III, 2009; Blitzer *et al.*, 2006], named-entity recognition and shallow parsing [Daumé III, 2009]. In general these algorithms identify a feature mapping where the source and target domains are similar. It includes learning non-linear mappings [Daumé III, 2009; Pan *et al.*, 2011], mappings to mitigate domain divergence [Pan *et al.*, 2010], common features [Dai *et al.*, 2007; Dhillon *et al.*, 2003], and ensemble based approaches [Bhatt *et al.*, 2015].

Multi-source domain adaptation has been explored in theory as well as in practical applications [Sun *et al.*, 2015]. Crammer *et al.* [2008], Mansoor *et al.*[2009], and Ben-David *et al.* [2010a] without proposing any algorithm, presented theoretical justifications for different methods on combining information from multiple sources. Empirical methods for multi-source adaptation can be broadly categorized into: 1) feature representation approaches including techniques to learn representations [Chattopadhyay *et al.*, 2012; Sun *et al.*, 2011; Duan *et al.*, 2009; L. Duan, 2012] to minimize divergence between domains, using marginal as well as conditional probability distributions across multiple domains or feature expansion during train and test using external information [Bollegala *et al.*, 2013]; 2) combining pre-trained classifiers including multiple convex combination of pre-learned classifiers by Schweikert *et al.* [2008], dynamic Bayesian learning framework by Sun and Shi [2013], adaptive support vector machines by Yang *et al.* [2007], multi-view adaboost transfer learning by Xu and Sun [Xu and Sun, 2012], and part based multi-view ensemble approach [Sun *et al.*, 2013]. While our approach also uses an ensemble-based combination of source classifiers, we provide an integrated algorithm which takes into account all the challenges listed in the introduction.

3 Preliminaries and Notation

Let $\mathcal{X}$ denote an instance space. A *domain* S is characterized by a probability distribution D_S over $\mathcal{X} \times \{+1, -1\}$. Let $D_S(A)$ denotes the probability of a set $A \subseteq \mathcal{X}$ under D_S. In our setting, the learner has access to a finite set of labeled instances from M different *source domains* and a finite set of unlabeled instances from a *target domain*. In particular, for each $k = 1$ to M, the learner has access to n_k i.i.d labeled instances drawn from the k-th source domain denoted by $\{\mathbf{x}_i^k, y_i^k\}_{i=1}^{n_k}$ and n_t unlabeled i.i.d instances drawn from the target domain denoted by $\{\mathbf{x}_i^T\}_{i=1}^{n_t}$. Let P_u and P_s be the pool of unlabeled and pseudo labeled instances from the target. The goal of the learner is to learn a classifier $C_T : \mathcal{X} \to \{+1, -1\}$ that predicts well in the target domain. Let C_k be a classifier trained on domain k and $\theta^k(\mathbf{x}) \in [-1, 1]$ be the *confidence* of C_k in predicting instance $\mathbf{x}$. Here, the sign denotes the class and magnitude denotes the confidence level. For instance, if C_k is a linear support vector machine corresponding to a hyper-plane $\mathbf{v}$, then $\theta^k(\mathbf{x}) = \frac{<\mathbf{v},\mathbf{x}>}{\|\mathbf{v}\|_2}$. Let $\ell(a, b) = (a - b)^2$ be the squared loss.

4 Source Selection

Theoretical results in domain adaptation literature [Ben-David *et al.*, 2010b] show that *adapting* to a target domain by learning a classifier using a source domain depends on a measure of similarity between the source and the target domains. We call this as $\mathcal{H}$-similarity [1] and define it as follows:

Definition 1 *Let $\mathcal{H} : \mathcal{X} \to \{1, -1\}$ be a hypothesis class. The $\mathcal{H}$-similarity between two domains S and T over $\mathcal{X}$ is defined as follows:*

$$Sim_{\mathcal{H}}(S, T) = 1 - \sup_{g \in \mathcal{H}\Delta\mathcal{H}} |D_S(\mathbf{I}_g) - D_T(\mathbf{I}_g)|$$

where $\mathbf{I}_g = \{x \in \mathcal{X} : g(x) = 1\}$ and $g \in \mathcal{H}\Delta\mathcal{H} \iff g(x) = h(x) \oplus h'(x)$ for some $h, h' \in \mathcal{H}$.

where $h(x) \oplus h'(x) = 1$ iff $h(x) \neq h'(x)$ and -1 otherwise. The $\mathcal{H}$-similarity depends on the worst case absolute difference in probabilities assigned to the regions where pairs of hypotheses from $\mathcal{H}$ disagree.

When the number of sources is large, one may not want to use all the sources for adaptation. This is critical as obtaining labels even for the source domains may be expensive/time consuming. Moreover, it is often computationally cheaper to work only with a few relevant sources. In such cases, one wishes to choose the best K sources (where $K \ll M$) suitable for adaptation. A natural choice is to order the source domains based on their $\mathcal{H}$-similarity to the target domain and pick the top K similar sources. An immediate question is whether this choice is *optimal* i.e., can there be a different combination of sources that has higher $\mathcal{H}$-similarity to the target domain? We show below, perhaps somewhat surprisingly that this natural choice may fail to be optimal.

Theorem 2 *There exists a set of M source domains, a target domain and a K such that any convex combination of the top K sources chosen based on the $\mathcal{H}$-similarity w.r.t the target domain is sub-optimal.*

Proof sketch: The key idea is to construct three domains $S_1 = \{\mathbf{x}_1\}, S_2 = \{\mathbf{x}_2\}, S_3 = \{\mathbf{x}_3\}$ each consisting of carefully chosen singletons such that a linear hypothesis class $\mathcal{H}$ cannot distinguish between two of the three domains i.e. $h(\mathbf{x}_1) = h(\mathbf{x}_2)\ \forall h \in \mathcal{H}$. Now if the target domain is more $\mathcal{H}$-similar to S_1 and S_2 than S_3, then choosing sources based only on individual similarity to the target will result in the choice $\{S_1, S_2\}$ for $K = 2$ whereas either $\{S_1, S_3\}$ or $\{S_2, S_3\}$ will be more $\mathcal{H}$-similar to the target than $\{S_1, S_2\}$.

[1] $\mathcal{H}$-similarity is a slightly modified version of the $\mathcal{H}$-divergence first defined in [Ben-David *et al.*, 2010b]. We define it as similarity here as opposed to divergence for the sake of clarity.

Algorithm 1 Greedy Algorithm for Selecting Sources

Input: A collection $\{S_1, S_2, \ldots S_M\}$ of source domains with respective pivot feature sets $P_1, P_2, \ldots P_M$, target T.

Obtain a permutation σ by sorting the source domains based on decreasing order of $\mathcal{H}$-similarity to T.
$S' = \{S_{\sigma(1)}\}$, $P' = P_{\sigma(1)}$, $K = 1$
Iterate: $j = 2$ to M
if $\text{Comp}(S_\sigma(j), S') \geq \frac{|P'|}{(K+1)!}$ **then**
$S' = S' \cup \{S_{\sigma(j)}\}$, $P' = P' \cup P_{\sigma(j)}$, $K = K + 1$,
end if.
end iterate.
Output: S', the set of K selected source domains.

The reason why choosing sources based on $\mathcal{H}$-similarity w.r.t the target is not optimal is because we have not taken into account how *complementary* the chosen source domains are to each other (Definition 3 in Section 4.1). In the following section, we propose an algorithm which uses this insight to pick the top K sources not only based on $\mathcal{H}$-similarity but also based on complementarity of sources.

4.1 Greedy Algorithm for Top K Source Selection

In this section, we use insights gained from Theorem 2 to propose a new algorithm for selecting the top K sources for a target domain, described in Algorithm 1. We emphasize that our algorithm does *not* need any labels and uses only unlabeled instances from both the source and the target domain to output the list of top K sources. It first orders the sources based on decreasing $\mathcal{H}$-similarity to the target. One can compute this by labeling the source domain instances as $+1$, target domain as -1 and computing the error of the best classifier that separates the source from the target [Ben-David *et al.*, 2010b]. Starting with the most similar source, the algorithm iteratively adds new sources to the list such that the selected sources are 'complementary' to each other.

Complemenetarity: We measure complementarity of a domain w.r.t another using *pivot* features. Pivot features are those features which are frequent and common in both the domains [Blitzer *et al.*, 2006; 2007]). Let P_i denote the set of pivot features corresponding to source i.

Definition 3 *The* complementarity *of domain S_i w.r.t a set of domains S' is given by*

$$\textit{Comp}(S_i, S') = |P_i - (P_i \cap P')| \tag{1}$$

where $P' = \cup_{j \in S'} P_j$

Algorithm 1 adds a source to the list of sources if complementarity of the source w.r.t to the list selected so far is above a threshold. In other words, the objective of the algorithm is to find the subset of sources $S' \subseteq [M]$ that maximizes both $\sum_{S_i \in S'} \text{sim}_{\mathcal{H}}(S_i, T)$ and $|\cup_{i \in S'} P_i|$.

The above problem can be cast as a maximum coverage problem for which finding an optimal S' is in general NP-hard [Vazirani, 2003]. We note that Algorithm 1 can be viewed as a modified version of the greedy algorithm for the maximum coverage problem.

5 Multi-source Iterative Domain Adaptation

In this section, we propose the multi source iterative domain adaptation (MSIDA) algorithm which is designed to satisfy two key objectives: 1) adapt to learn target specific features building on the generalized representations, and 2) gradually transfer knowledge from multiple sources. MSIDA algorithm builds on the K source domains selected using Algorithm 1 and its different stages are explained in the next sections.

5.1 Adapting Beyond Generalization

A *good* common shared representation [Blitzer *et al.*, 2007; Ji *et al.*, 2011; Pan *et al.*, 2010] mitigates divergence between source and target domains and thus, the cross-domain classification error. The proposed algorithm, MSIDA, builds on one such widely used common representation, Structural Correspondence Learning (SCL) [Blitzer *et al.*, 2007; 2006], which starts by identifying frequent common features between a source-target pair as *pivots*. A linear predictor is trained for each pivot feature whose weights constitute a column vector in matrix W. This matrix captures the co-occurrence between features expressing similar meaning in different domains. Top k Eigenvectors of matrix W represent the principal predictors for the weight space, Q. Features from both domains are projected on this principal predictor space, Q, to obtain a shared representation. Readers are kindly referred to [Blitzer *et al.*, 2006; 2007] for further details about SCL.

A classifier trained using labeled data from the source domain, on this shared representation, is expected to generalize well on the target domain. We extend this principle of "generalization" to multi-source adaption and train classifier C_k on shared representation of k^{th} source-target pair, referred to as "source classifier". Further to adapt to the target domain, unlabeled target instances (projected on the shared representations) are passed through all source classifiers and their outputs are suitably combined to obtain predictions. Confidently predicted instances can be considered as pseudo labeled data and a classifier, C_T, can be initialized using target specific representation. C_T captures a different view of target domain, than the shared representations, to include domain specific features and thus referred to as "target classifier".

5.2 Gradual Iterative Adaptation

Only a handful of instances in target domain are confidently predicted using the shared representations from multiple source domains, therefore, we further iterate to create additional pseudo labeled instances in target domain. In next round of iterations, remaining unlabeled instances are passed through a weighted ensemble of the target and all source classifiers with their outputs combined as shown below:

$$E(\mathbf{x}^T) = \hat{y} = \text{sign}\left(w^T \cdot \theta^T(\mathbf{x}^T) + \sum_{k=1}^{K} w^k \cdot \theta^k(Q_k \mathbf{x}^T) \right) \tag{2}$$

where $\hat{y}$ is the predicted label. In j^{th} iteration, let P_j be the number of confidently labeled instances added to the pool of pseudo labeled data (P_s) using which classifier in target domain is updated/re-trained. This process is repeated till all

Algorithm 2 Multi-source Iterative Learning Algorithm

Input: K source domains, P_u unlabeled instances in target, thresholds θ_1, θ_2.
Initialize $P_s = \emptyset$
$\forall k$, estimate Q_k and train C_k on $\{Q_k \mathbf{x}_i^k, y_i^k\}_{i=1}^{n_k}$
for $i = 1$ **to** n_t **do**
  $\alpha_i = \sum_{k=1}^{K} \theta^k(\mathbf{x}_i^T)$; $\hat{y}_i = \text{sign}(\alpha_i)$
  if $\alpha_i > \theta_1$ **then**
    Move $\mathbf{x}_i^T$ from P_u to P_s with pseudo label $\hat{y}_i$.
  end if.
end for.

Initialize $\forall k, w_0^k = w_0^T = \frac{1}{K+1}$; Train C_T on P_s
for $j = 0$: till $P_u = \emptyset$ or $j \leq iterMax$ **do**
  for $i = 1$ to $|P_u|$ **do**

$$E(\mathbf{x}_i^T) = \text{sign}\left(w_j^T \theta^T(\mathbf{x}_i^T) + \sum_{k=1}^{K} w_j^k \theta^k(Q_k \mathbf{x}_i^T) \right)$$

$$\alpha_i = \sum_{k=1}^{K} \theta^k(\mathbf{x}_i^T) + \theta^T(\mathbf{x}_i^T)$$

  if $\alpha_i > \theta_2$ **then**
    Move $\mathbf{x}_i^T$ from P_u to P_s with pseudo label $\hat{y}_i = E(\mathbf{x}_i^T)$.
  end if.
  end for.
  Retrain C_T on P_s and update all w^k and w^T according to Equations 3 and 4.
end for
Output: Updated C_T, w_j^k and w_j^T.

unlabeled data is labeled or certain maximum number of iterations is performed. The gradual transfer of knowledge occurs within the ensemble as the source classifiers contribute in transforming unlabeled instances to pseudo labeled instances to append to the training of target classifier. Further to mitigate the effects of negative transfer, which is generally experienced during aggressive one-shot transfer, ensemble weights are also progressively updated based on the loss of individual classifiers in each iteration as shown below:

$$w_{(j+1)}^k = \frac{w_j^k \cdot I(C_k)}{\sum_{k=1}^{K} w_j^k \cdot I(C_k) + w_j^T \cdot I(C_T)} \tag{3}$$

$$w_{(j+1)}^T = \frac{w_j^T \cdot I(C_T)}{\sum_{k=1}^{K} w_j^k \cdot I(C_S) + w_j^T \cdot I(C_T)} \tag{4}$$

$I(\cdot)$ is loss function to measure the error of individual classifiers w.r.t. to the pseudo labels obtained by the ensemble over all confidently predicted instances in an iteration:

$$I(C_k) = \exp\left(-\eta \sum_{p=1}^{P_j} \ell(\theta^k(\mathbf{x}_p^T), \hat{y_p}) \right) \tag{5}$$

Subsequently, this learned ensemble is used for classification in the target domain. Finally, Algorithm 2 summarizes the proposed multi-source iterative domain adaptation algorithm.

5.3 Error Bounds

We now analyze the mistake bound of MSIDA and introduce the following proposition.

Proposition 4 *Using the exponential weighting update as in Equation 3 and 4 and setting $\eta = 0.5$ in Algorithm 2, we have*

$$\sum_{i=1}^{n_t} \ell\left(\sum_{k=1}^{K} w_i^k \theta^k(\mathbf{x}_i^T) + w_i^t \theta^T(\mathbf{x}_i^T), \hat{y}_i \right) \leq \frac{\ln(K+1)}{\eta} + \min\left\{ \sum_{i=1}^{n_t} \ell\left(\sum_{k=1}^{K} \theta^k(\mathbf{x}_i^T), \hat{y}_i \right), \sum_{i=1}^{n_t} \ell\left(\theta^T(\mathbf{x}_i^T), \hat{y}_i \right) \right\}$$

Theorem 5 *Let $A_k = \sum_{i=1}^{n_t} \ell\left(\theta^k(\mathbf{x}_i^T), \hat{y}_i\right), B = \sum_{i=1}^{n_t} \ell\left(\theta^T(\mathbf{x}_i^T), \hat{y}_i\right)$ be the loss of the k-th source classifier and the target classifier respectively. Then, the number of mistakes M made by the MSIDA algorithm is bounded as*

$$M \leq 4 \min\left\{ \sum_{k=1}^{K} A_k, B \right\} + 8 \ln(K+1)$$

Proof sketch: When there is a mistake at some i^{th} instance, we have $\left| (\sum_{k=1}^{K} w_i^k \theta^k + w_i^T \theta^T - \hat{y} \right| \geq \frac{1}{2}$. Thus, we have

$$\sum_{i=1}^{n_t} \ell\left(\sum_{k=1}^{K} w_i^k \theta^k + w_i^t \theta^T, \hat{y}_i \right) = \sum_{i=1}^{n_t} \left(\left(\sum_{k=1}^{K} w_i^k \theta^k + w_i^t \theta^T \right) - \hat{y}_i \right)^2 \geq \frac{1}{4} M \tag{6}$$

Combining Eq 6 with Proposition 4 (following the proof in [Zhao and Hoi, 2010]), we have:

$$\frac{1}{4} M \leq \min\left\{ \sum_{k=1}^{K} \sum \theta^k, \sum \theta^T \right\} + \frac{ln(K+1)}{\eta} \tag{7}$$

Multiplying the inequality in Eq. 7 by 4 gives Theorem 5. Note that $l(\theta^k(x_i^T), \hat{y}_i)$ is an upper bound of $\frac{1}{4} A_k$ instead of A_k (because l is a square loss and both $\theta^k(x_i^T)$ and $\hat{y}_i$ are normalized to $[0, 1]$); similarly, $l(\theta^T(x_i^T), \hat{y}_i)$ is the upper bound of $\frac{1}{4} B$. Further, if we assume $l(\theta^k(x_i^T), \hat{y}_i) \sim \frac{1}{4} A_k$ and $l(\theta^T(x_i^T), \hat{y}_i) \sim \frac{1}{4} B$, we have: $M \leq 4 \min\{\sum_{k=1}^{K} \{A_k, B\}\} + 8ln(K+1)$. This gives a strong theoretical support for the MSIDA algorithm.

6 Experiments

The efficacy of the proposed algorithm is evaluated for cross-domain sentiment classification task and the performance is reported in terms of classification accuracy.

6.1 Datasets

The first dataset comprises the widely used Amazon review dataset [Blitzer *et al.*, 2007] appended with the Amazon product dataset [McAuley *et al.*, 2015b; 2015a] for evaluating the

Table 1: Target collections from the OSM dataset used.

Target	Description	Unlabeled	Test
Coll1	Mobile support	22645	650
Coll2	Obama Healthcare	36902	1050
Coll3	Microsoft kinnect	20907	258
Coll4	X-box	36000	580

Table 2: Lists the product domains from the Amazon product dataset that are used as existing source domains.

Domains	
Clothing, Shoes & Jewelry	CD & Vinyl
Kindle Store	Sports and Outdoors
Cell Phones & Accessories	Health & Personal Care
Toys & Games	Video Games
Tools & Home Improvement	Beauty
Apps for Android	Office Products
Pet Supplies	Automotive
Grocery & Gourmet Food	Patio, Lawn and Garden
Baby	Digital Music
Musical Instruments	Amazon Instant Video

challenges of multi-source adaptation. The Amazon review dataset has four domains, Books(*B*), Dvds(*D*), Kitchen(*K*) and Electronics(*E*) which serve as target domains. The Amazon product dataset comprises 20 additional domains, as shown in Table 2, to serve as source domains. Each domain comprises 2000 (equal positive and negative) reviews. 1600 reviews from each source-target domain pair are used to learn shared representations. For source domains, these reviews are also used to learn the source classifiers. For target domains, 1600 reviews serve as unlabeled pool of instances and the performance is reported on the non-overlapping 400 reviews.

The second dataset is from Twitter.com and is referred to as online social media (OSM) dataset. It has 75 collections, each comprising tweets about products and services from different domains such as telecom, retail, finance, health care and hotel & transportation *etc.* We want to emphasize that even collections from the same domains (e.g. Apple iPhone and Samsung S3) may vary quite significantly with respect to features and label distributions; hence are considered as separate domains for uniform nomenclature. As shown in Table 1, we randomly selected 4 (out of 75) collections to serve as target collections and the remaining collections as source collections. For each source collection, we used 10000 tweets to learn shared representations and source classifiers.

6.2 Experimental Protocol

The performance of the proposed algorithm, referred to as **MSIDA**, is compared with different techniques. A baseline algorithm (**BL**) where a classifier trained on a single (most similar) source domain is directly applied on the target domain and an in-domain classifier (**GOLD**) which is trained and tested on the data from the target domain. We also compared the performance of the proposed algorithm with existing domain adaptation approaches including structural correspondence learning (**SCL**) [Blitzer, 2007; Blitzer *et al.*, 2006] , spectral feature alignment (**SFA**) [Pan *et al.*, 2010], and a recently proposed multi-source adaptation algorithm using sentiment thesaurus (**ST**) [Bollegala *et al.*, 2013]. In our experiments, we used SVM classifiers with

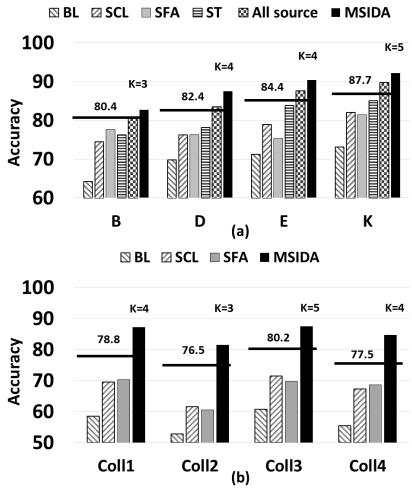

Figure 2: Compares the performance of MSIDA with existing techniques on the (a) Amazon review and (b) OSM datasets.

radial basis function (RBF) kernels as individual classifiers combined with uniformly initialized weights in the ensemble and the maximum number of iterations ($iterMax$) set to 30.

6.3 Results and Analysis

Cross-domain Sentiment Classification

Figure 2 shows that the proposed "MSIDA" algorithm significantly outperforms existing algorithms on the two datasets. It also exceeds the in-domain (**GOLD**) performance (horizontal lines on the plot) by leveraging rich complementary information from multiple source domains. For the Amazon dataset, we first consider only 4 domains, namely books(*B*), electronics(*E*), dvds(*D*) and kitchen(*K*). For a selected target domain, the proposed algorithm considers the remaining 3 domains as the source domains, referred to as "All source", a widely used setting in literature for multi-source adaptation on the Amazon review dataset. In the second case, we consider all domains listed in Table 2 as candidate source domains and select K optimal source domains, referred to as "MSIDA". Further at 95% confidence interval, parametric t-tests suggest that the "MSIDA" algorithm is statistically significantly different from existing approaches such as SCL and SFA.

Effect of Multiple Sources

Results in Figure 3 show that the performance of the proposed MSIDA algorithm exceeds the performance when adapting from single most similar source by 5% and 14% (on average) on the two datasets respectively. The proposed algorithm also exceeds the adaptation performance with random K source domains by 16% and 19% and (on average) on the two datasets respectively. Adapting from random K source domains was observed to be inferior to adapting from the single most similar domain as the former does not maximizes either similarity with the target domain or complementary information among source domains.

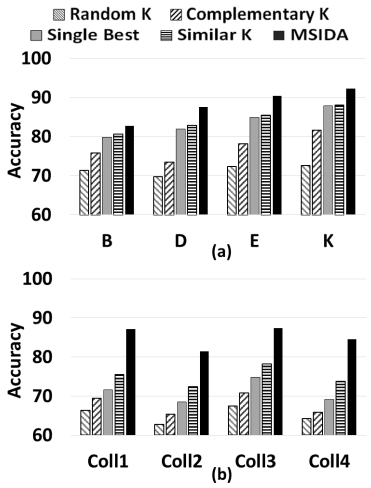

Figure 3: Compares different approaches to combine multiple sources on the (a) Amazon review and (b) OSM datasets.

Table 3: Performances of different stages of MSIDA.

Dataset	Target	All source classifiers	C_T at $l = 0$	E	All source data	MSIDA
Amazon review dataset	B	64.2	58.5	68.4	61.8	**82.7**
	D	69.8	60.4	70.3	65.4	**87.5**
	E	71.3	64.9	75.4	67.2	**90.4**
	K	73.1	67.5	77.2	67.8	**92.2**
OSM dataset	Coll1	58.6	40.5	61.7	54.6	**87.2**
	Coll2	52.8	38.4	56.2	49.4	**81.5**
	Coll3	60.8	50.8	64.6	55.2	**87.4**
	Coll4	55.4	40.7	60.8	50.5	**84.6**

Effect of Similarity and Complementariness

We compared the performance of the MSIDA algorithm with adaptation from top-K similar and top-K complementary source domains, referred to as "Similar K" and "Complementary K" respectively. Figure 3 shows that the performance of the MSIDA algorithm exceeds "Similar K" by 4% and 10% (on average) and also outperforms "Complementary K" by 10% and 17% (on average) on the two datasets respectively. Overall analysis suggests that simultaneously maximizing similarity and complementariness lead to significant improvements in the performance.

Effect of Iterative Learning

Results in Table 3 show that the proposed algorithm outperforms the combination of all source classifiers trained on shared representations ("All source classifiers") by at least 20% & 32% and an algorithm trained on data combined from K source domains ("All source data") by at least 17% \$ 26% on the two datasets respectively. It validates that learning target specific features enhances the target performance.

Effect of varying threshold θ_1 & θ_2

C_t may be trained on incorrect pseudo labeled instances if θ_1 is low; whereas, if θ_1 is high, C_t may be deficient of instances to learn good decision boundary. On the other hand, when θ_2 is low, the algorithm does not benefits from the iterative learning; whereas, a high θ_2 tends to make the algorithm starve for

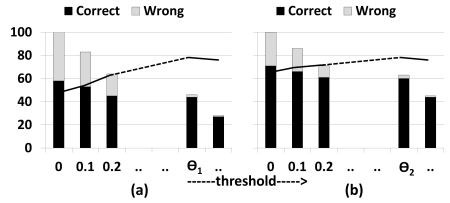

Figure 4: Bar plot shows % of unlabeled data that crosses confidence threshold, lower and upper part of the bar represents % correctly and wrongly predicted pseudo labels.

Table 4: Comparing different shared representations.

Dataset	Target	Common	MVPCA	SFA	SCL
Amazon review dataset	B	71.4	76.1	80.1	**82.7**
	D	73.8	77.5	84.0	**87.5**
	E	78.6	82.8	86.5	**90.4**
	K	79.8	83.5	89.8	**92.2**
OSM dataset	Coll1	66.8	76.4	84.8	**87.2**
	Coll2	71.0	79.2	79.8	**81.5**
	Coll3	74.4	81.2	86.2	**87.4**
	Coll4	73.5	79.4	83.3	84.6

pseudo labeled instances. θ_1 and θ_2 are set empirically on a held-out set and the *knee-shaped* curve in Figure4 shows that there exists an optimal value for both θ_1 and θ_2.

Effect of different shared representations

The performance is compared when MSIDA algorithm uses different shared representations than SCL. We used 1) common features between two domains ("common"), 2) multi-view principal component analysis based representation ("MVPCA") [Ji *et al.*, 2011] and 3) spectral feature alignment ("SFA") [Pan *et al.*, 2010] as these have been previously used in literature. Table 4 indicates that the proposed algorithm performs well with any representation that captures commonalities between the source and target domains.

7 Conclusions

This paper proposed a novel multi-source iterative domain adaptation (MSIDA) algorithm that leverages knowledge from multiple sources to learn an accurate classifier for the target domain. Its first contribution is a novel technique to select representative source domains from multiple sources by simultaneously maximizing the similarity to the target as well as the complementary information among the selected sources. Its second contribution is an ensemble based iterative adaptation algorithm that progressively learns domain specific features from the unlabeled target domain data starting with a shared feature representation. In each iteration, pseudo labeled instances are used to update the target classifier and the contributions of individual classifiers in the ensemble. Finally, for cross-domain classification on the real world and standard benchmark datasets, we demonstrated the efficacy of the proposed algorithm where it significantly outperformed all existing approaches.

References

[Ando and Zhang, 2005] Rie Kubota Ando and Tong Zhang. A high-performance semi-supervised learning method for text chunking. In *Proceedings of Association for Computational Linguistics*, pages 1–9, 2005.

[Ben-David *et al.*, 2010a] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

[Ben-David *et al.*, 2010b] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

[Bhatt *et al.*, 2015] H. S. Bhatt, D. Semwal, and S. Roy. An iterative similarity based adaptation technique for cross-domain text classification. In *Proceedings of Conference on Natural Language Learning*, pages 52–61, 2015.

[Blitzer *et al.*, 2006] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 120–128, 2006.

[Blitzer *et al.*, 2007] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of Association for Computational Linguistics*, pages 187–205, 2007.

[Blitzer, 2007] J Blitzer. Domain adaptation of natural language processing systems. Technical Report PhD Thesis, The University of Pennsylvania, 2007.

[Bollegala *et al.*, 2013] D. Bollegala, D. Weir, and J. Carroll. Cross-domain sentiment classification using a sentiment sensitive thesaurus. *IEEE Transactions on Knowledge and Data Engineering*, 25(8):1719–1731, 2013.

[Chattopadhyay *et al.*, 2012] Rita Chattopadhyay, Qian Sun, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Multisource domain adaptation and its application to early detection of fatigue. *Transactions on Knowledge Discovery and Databases*, 6(4):1–26, 2012.

[Chen *et al.*, 2009] Bo Chen, Wai Lam, Ivor Tsang, and Tak-Lam Wong. Extracting discriminative concepts for domain adaptation in text mining. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 179–188, 2009.

[Dai *et al.*, 2007] W Dai, G-R Xue, Q Yang, and Y Yu. Co-clustering based classification for out-of-domain documents. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 210–219, 2007.

[Daumé III, 2009] Hal Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.

[Dhillon *et al.*, 2003] I. S. Dhillon, S. Mallela, and D. S Modha. Information-theoretic co-clustering. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 89–98, 2003.

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

[Ji *et al.*, 2011] Y. S. Ji, J. J. Chen, G. Niu, L. Shang, and X. Y. Dai. Transfer learning via multi-view principal component analysis. *Journal of Computer Science and Technology*, 26(1):81–98, 2011.

[K. Crammer, 2008] J. Wortman. K. Crammer, M. Kearns. Learning from multiple sources. *Journal of Machine Learning Research*, 9(1):17571774, 2008.

[L. Duan, 2012] I. Tsang L. Duan, D. Xu. Domain adaptation from multiple sources: a domaindependent regularization approach. *IEEE Transactions on Neural Networks and Learning Systems*, 23(3):504518, 2012.

[McAuley *et al.*, 2015a] Julian McAuley, Rahul Pandey, and Jure Leskovec. Inferring networks of substitutable and complementary products. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2015.

[McAuley *et al.*, 2015b] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52, 2015.

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

[Pan *et al.*, 2010] Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of International Conference on World Wide Web*, pages 751–760, 2010.

[Pan *et al.*, 2011] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

[Schweikert and Widmer, 2008] Gabriele Schweikert and Christian Widmer. An empirical analysis of domain adaptation algorithms for genomic sequence analysis. In *Advances in Neural Information Processing Systems*, pages 1433–1440, 2008.

[Sun and Shi, 2013] S. Sun and H. Shi. Bayesian multi-source domain adaptations. In *International Conference on Machine Learning and Cybernetics*, pages 24–28, 2013.

[Sun *et al.*, 2011] Qian Sun, Rita Chattopadhyay, Sethuraman Panchanathan, and Jieping Ye. A two-stage weighting framework for multi-source domain adaptation. In *Advances in Neural Information Processing Systems*, pages 505–513. 2011.

[Sun *et al.*, 2013] Shiliang Sun, Zhijie Xu, and Mo Yang. Transfer learning with part-based ensembles. In *Multiple Classifier Systems*, volume 7872, pages 271–282. 2013.

[Sun *et al.*, 2015] Shiliang Sun, Honglei Shi, and Yuanbin Wu. A survey of multi-source domain adaptation. *Information Fusion*, 24:84 – 92, 2015.

[Vazirani, 2003] Vijay V. Vazirani. *Approximation Algorithms*. Springer-Verlag Berlin Heidelberg, 2003.

[Xu and Sun, 2012] Zhijie Xu and Shiliang Sun. Multi-source transfer learning with multi-view adaboost. In *Proceedings of International Conference on Neural Information Processing*, pages 332–339, 2012.

[Y. Mansour, 2009] A. Rostamizadeh Y. Mansour, M. Mohri. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems*, pages 1041–1048. 2009.

[Yang *et al.*, 2007] Jun Yang, Rong Yan, and Alexander G. Hauptmann. Cross-domain video concept detection using adaptive svms. In *Proceedings of International Conference on Multimedia*, pages 188–197, 2007.

[Zhao and Hoi, 2010] P. Zhao and S. Hoi. OTL: A framework of online transfer learning. In *Proceedings of International Conference on Machine Learning*, pages 1231–1238, 2010.