Source Free Domain Adaptation Using an Off-the-Shelf Classifier

Arun Reddy Nelakurthi
Arizona State University
Tempe, USA
anelakur@asu.edu

Ross Maciejewski
Arizona State University
Tempe, USA
rmacieje@asu.edu

Jingrui He
Arizona State University
Tempe, USA
jingrui.he@asu.edu

***Abstract*—With the advancements in many data mining and machine learning tasks, together with the availability of large-scale annotated data sets, there have been an increasing number of off-the-shelf tools for addressing these tasks, like Stanford NLP Toolkit and Caffe Model Zoo. However, many of these tasks are time-evolving in nature due to, e.g., the emergence of new features and the change of class conditional distribution of features. As a result, the off-the-shelf tools are not able to adapt to such changes and will suffer from sub-optimal performance in the target application. In this paper, we propose a generic framework named AOT for adapting the outputs from an off-the-shelf tool to accommodate the changes in the learning task. It considers two major types of changes, i.e., label deficiency and distribution shift, and aims to maximally boost the performance of the off-the-shelf tool in the target domain, with the help of a limited number of target domain labeled examples. Furthermore, we propose an iterative algorithm to solve the resulting optimization problem, and we demonstrate the superior performance of the proposed AOT framework on text and image data sets.**

***Index Terms*—off-the-shelf classifiers, label deficiency, distribution shift**

I. Introduction

In the past decades, the advancements in the field of machine learning have led to their wide adoption to solve different real world applications. In general, training a new machine learning model needs large amount of labeled data. In some applications, such large-scale annotated data sets are readily available, giving rise to an increasing number of off-the-shelf tools For example, the *Caffe Model Zoo* [1] hosts different models that can be readily used for various classification tasks; language processing tools such as *Stanford NLP Toolkit* [2] come with various models for natural language processing tasks. However, many of these machine learning tasks are time-evolving in nature due to, e.g., the emergence of new features and the shift in class conditional distribution. As a result, the off-the-shelf tools may not be able to adapt to such changes in a timely fashion, and will suffer from sub-optimal performance in the learning task.

On the other hand, existing work on transfer learning, cannot be readily applied to improve the performance of the off-the-shelf tools due to the lack of the training data for obtaining these tools, i.e., the lack of source domain data. More specifically, due to licensing or other copyright restrictions, the labeled data sets sometimes are not released but the underlying models are made available to use as a black-box classifier [3]. Therefore, given these black box classifiers, *is it possible to leverage these classifiers to improve the classification performance on the evolved source domain, i.e., the target domain, given limited amount of training data from the target domain?* To this end, we are facing two major challenges: (1) Label deficiency happens when new features appear, or the relationship between individual features and the class labels changes in the target domain; (2) Distribution shift happens when the class conditional distribution in the target domain is different from the training data used by off-the-shelf classifiers, potentially changing the optimal predicted label.

In this paper we address the above mentioned challenges of label deficiency and distribution shift through the proposed **A**daptive **O**ff-**T**he-shelf classification (**AOT**) framework. In our case, we consider that there exists a black-box classifier that gives out the classification labels for the target domain examples, and no other information about the black-box is known. In particular, we assume that the training data used to obtain the off-the-shelf classifiers are not available, and we aim to leverage the noisy class labels predicted by the black-box classifier and very few labeled examples from the target domain to improve the classification performance of the unlabeled data in the target domain. Given an unlabeled document from the target domain, these tools are able to predict the polarity of the text without taking into consideration the unique characteristics of this domain. The proposed framework is able to effectively integrate the information from these tools as well as the few labeled examples from the target domain to construct a classification model for the target domain with significantly improved performance.

The following are the main contributions of our paper; (1) A novel problem setting of source free domain adaptation, where the goal is to leverage the output of an off-the-shelf classifier and a few labeled examples from the target domain, in order to obtain a significantly better classification model for the target domain, as compared to the off-the-shelf classifier; (2) A generic optimization framework named **AOT** to adapt an off-the-shelf classifier to the target domain by explicitly addressing the two major types of changes from the source domain to the target domain, i.e., label deficiency and distribution shift; (3) Analysis on the performance of the proposed **AOT** framework in terms of convergence to the global optimum, and the complexity of the proposed algorithm.

II. Related Work

Transfer learning is a widely studied problem. Different supervised, unsupervised and semi-supervised methods have been proposed for a wide variety of applications such as machine translation [4], image classification [5] and web document classification [6]. Source free transfer learning is a special case of transfer learning, with limited access to source domain data. Often, it is considered that either the class distribution of the source domain data is known or there is limited access to the source domain data in the form of representative examples. In [7], [8], authors proposed an Adaptive-SVM framework called DAM where the goal is to learn the target classification function by adapting the pre-trained classifiers to the labeled examples in the target domain. Our work is significantly different from DAM, as we consider only one off-the-shelf classifier compared to multiple SVM classifiers used in DAM and also provide a drift correction framework to adapt the off-the-shelf classifier to labeled examples. Also, unlike DAM, our generic framework works with any off-the-shelf classifier, not restricted to kernelized SVM. The authors of [9] proposed a source domain free approach by leveraging the information from existing knowledge sources such as WWW or Wikipedia to compute the target labels from unlabeled examples. The problem of 'source free transfer learning' in [9] is different from the problem setting studied in this paper; instead of building a knowledge base we simply make use of an existing off-the-shelf black-box classifier to improve prediction accuracy on the set of unlabeled examples in the target domain. In [10], the authors consider three different scenarios: (1) the parameters of the source classifiers are known; (2) source classifiers as a black box; (3) class distribution of the source classifier is known. Scenario 2 is very relevant to our work where they employed marginalized denoising autoencoders to denoise the source classifier labels using unlabeled data in the target domain. Notice that this work assumes that the difference between source and target domain classifiers is linear with respect to the features of target domain examples, whereas we allow this difference to take more complex forms, accommodating various degree of task relatedness. In addition, compared with all the existing work in this direction, we explicitly address the two major types of changes f rom the source domain to the target domain, namely label deficiency and distribution shift, whereas existing work only considers one of them.

III. Problem Definition

In this section, we introduce the notation used in the paper and formally define the problem of source free domain adaptation with an off-the-shelf classifier. Let $\mathcal{D}_L = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$ be the set of m labeled examples from the target domain; $\mathcal{D}_U = \{(\mathbf{x}_i)\}_{i=m+1}^{m+n}$ be the set of n unlabeled examples from the target domain, where $\mathbf{x}_i \in \mathbb{R}^d$ is a real valued vector of size d; and $y_i \in \{-1, 1\}, \forall i \in 1, \ldots, m$ is the binary class label. We consider the number of labeled examples to be much smaller than the number of unlabeled examples, i.e., $m \ll n$. Let $\mathbf{f}^0 = [y_1^0, \ldots, y_{m+n}^0]^T$ be a $(m+n)$-dimensional vector consisting of the pseudo-labels generated by the off-the-shelf classifier, where $y_i^0 \in \{-1, 1\}, i \in 1, \ldots, m+n$, and $c_i \in [0, 1)$, $i = 1, \ldots, m+n$, be the confidence score for each of the $m+n$ examples ($\mathcal{D}_L \cup \mathcal{D}_U$).

First of all, we represent all $m+n$ examples from the target domain as a graph $G = (V, E)$, where V is the set of nodes, and E is the set of edges. In this graph, each node corresponds to an example, labeled or unlabeled, i.e., $|V| = m+n$, and the weight associated with each edge measures the similarity between a pair of nodes. Let $\mathbf{W}$ be the affinity matrix of this graph, whose non-negative element $\mathbf{W}_{ij}$ in the i^{th} row and j^{th} column is the weight of the edge connecting the examples $\mathbf{x}_i$ and $\mathbf{x}_j$. Let $\mathbf{D}$ be the $(m+n) \times (m+n)$ degree matrix whose diagonal elements are set to be $\sum_j \mathbf{W}_{ij}$. The normalized Laplacian of the affinity matrix $\mathbf{W}$ is given by $\mathbf{S} = \mathbf{D}^{-\frac{1}{2}} \mathbf{W} \mathbf{D}^{-\frac{1}{2}}$.

The problem of source free domain adaptation is to adapt the noisy pseudo-labels $\mathbf{f}^0$ from the off-the-shelf classifier to the examples in the target domain by leveraging the information of a small number of labeled examples $\mathcal{D}_L$ from the target domain, without having access to the source domain data based on which the off-the-shelf classifier was trained. More specifically, given a set of m labeled examples $\mathcal{D}_L$; the n unlabeled examples $\mathcal{D}_U$; the normalized affinity matrix for the $m+n$ examples $\mathbf{S}$; and the noisy pseudo-labels $\mathbf{f}^0$ from the off-the-shelf classifier; the goal of source free domain adaptation is to learn a classification vector $\mathbf{f} \in \mathbb{R}^{m+n}$ to correctly classify all the $m+n$ examples. Notice that unlike $\mathbf{f}^0$, the elements of $\mathbf{f}$ may not be binary. Therefore, the predicted class label $\hat{\mathbf{y}}_i$ of the unlabeled examples $\mathcal{D}_U$ is set as $\hat{\mathbf{y}}_i = +1$, if $f_i \geq 0, i \in m+1, \ldots, m+n$, and $\hat{\mathbf{y}}_i = -1$ otherwise, where f_i is the i^{th} element of $\mathbf{f}$.

IV. The Proposed **AOT** Framework

In this section, we propose our **AOT** framework. The goal of the **AOT** framework is to learn the classification vector $\mathbf{f}$ for all the $m+n$ examples based on $\mathbf{f}^0$. Usually the pseudo-labels from the black-box classifier $\mathbf{f}^0$ are noisy due to label deficiency and distribution shift. As shown in eq. (1), we decompose the classification vector $\mathbf{f}$ into the sum of noisy pseudo-labels from off-the-shelf classifier $\mathbf{f}^0$ and two residual vectors:

$$\mathbf{f} = \mathbf{f}^0 + \Delta_1 \mathbf{f} + \Delta_2 \mathbf{f} \tag{1}$$

where $\Delta_1 \mathbf{f} \in \mathbb{R}^{m+n}$ and $\Delta_2 \mathbf{f} \in \mathbb{R}^{m+n}$ are the residual vectors that address label deficiency and distribution shift respectively. More specifically, the residual vector $\Delta_1 \mathbf{f}$ accounts for the change in the relationship between features and class labels in the target domain. For example, in sentiment classification, with the emergence of new words in the target domain, $\Delta_1 \mathbf{f}$ will provide correcting information regarding the relationship between the new words and the class labels. On the other hand, the residual vector $\Delta_2 \mathbf{f}$ addresses the changes in class conditional distribution in the target domain compared to the source domain data used to train the off-the-shelf classifier. For example, in sentiment classification, $\Delta_2 \mathbf{f}$ will provide

```
Algorithm 1: AOT : Adaptive Off-the-shelf Classifier
Input: (1) The normalized affinity matrix $\mathbf{S}$ for the $m + n$ examples in the
       target domain; (2) The noisy class labels $\mathbf{f}^0$ generated by the
       off-the-shelf classifier; (3) The max number of iterations $T$
Output: $\mathbf{f}$: The classification vector for all the examples in the target domain.
1 Initialize $\Delta\mathbf{f}_0 = \mathbf{0}^{m+n}$
2 for $t = 1$ to $T$ do
3     Fix $\Delta_1\mathbf{f}$, compute $\Delta_2\mathbf{f}$ using AddressLabDef
4     Fix $\Delta_2\mathbf{f}$, compute $\Delta_1\mathbf{f}$ using AddressDisShift
5 end
6 return $\mathbf{f} = \mathbf{f}^0 + \Delta_1\mathbf{f} + \Delta_2\mathbf{f}$
```

insights regarding the potentially different sentiment polarity for certain combination of keywords that are specific to the target domain.

In our framework, we propose to solve for both residual vectors via the following generic optimization problem.

$$Q(\Delta_1\mathbf{f}, \Delta_2\mathbf{f}) = Q_1(\Delta_1\mathbf{f}, \Delta_2\mathbf{f}) + Q_2(\Delta_1\mathbf{f}) + Q_3(\Delta_2\mathbf{f}) \quad (2)$$

where Q_1 takes into consideration both residual vectors, and it aims to enforce label consistency on all the examples along the data manifold; Q_2 is a sparsity constraint on the label deficiency residual vector $\Delta_1\mathbf{f}$; and Q_3 is the objective function of $\Delta_2\mathbf{f}$ for addressing the distribution shift. The optimal residual vectors $(\Delta_1\mathbf{f}^*, \Delta_2\mathbf{f}^*)$ is computed as follows:

$$(\Delta_1\mathbf{f}^*, \Delta_2\mathbf{f}^*) = \underset{\Delta_1\mathbf{f}\in\mathbb{R}^{m+n}, \Delta_2\mathbf{f}\in\mathbb{R}^{m+n}}{\operatorname{argmin}} Q(\Delta_1\mathbf{f}, \Delta_2\mathbf{f}) \quad (3)$$

To solve this optimization problem, we propose to use the alternating minimization strategy. More specifically:

$$\Delta_1\mathbf{f}_{t+1} = \underset{\Delta_1\mathbf{f}\in\mathbb{R}^{m+n}}{\operatorname{argmin}} Q_1(\Delta_1\mathbf{f}, \Delta_2\mathbf{f}_t) + Q_2(\Delta_1\mathbf{f}) \quad (4)$$

$$\Delta_2\mathbf{f}_{t+1} = \underset{\Delta_2\mathbf{f}\in\mathbb{R}^{m+n}}{\operatorname{argmin}} Q_1(\Delta_1\mathbf{f}_t, \Delta_2\mathbf{f}) + Q_3(\Delta_2\mathbf{f}) \quad (5)$$

where $t = 0, \ldots, T-1$, T is the total number of iterations, and $\Delta_1\mathbf{f}_t$ ($\Delta_2\mathbf{f}_t$) is the vector $\Delta_1\mathbf{f}$ ($\Delta_2\mathbf{f}$) in the t^{th} iteration. The proposed **AOT** algorithm (Algo. 1) runs till convergence or the max number of iterations is reached. It takes the normalized affinity matrix $\mathbf{S}$ and the set of noisy class labels generated by the off-the-shelf classifier $\mathbf{f}^0$ as input, and outputs the classification vector $\mathbf{f}$ for the examples in the target domain. After the initialization step (Step 1), the algorithm iteratively updates $\Delta_1\mathbf{f}$ and $\Delta_2\mathbf{f}$ in Steps 3 and 4. After T iterations or convergence, the **AOT** algorithm outputs the vector $\mathbf{f}$ for all the examples in the target domain. As discussed earlier, the predicted class label $\hat{\mathbf{y}}_i$ of the unlabeled examples $\mathcal{D}_U$ is set as $\hat{\mathbf{y}}_i = +1$, if $f_i \geq 0, i \in m+1, \ldots, m+n$, and $\hat{\mathbf{y}}_i = -1$ otherwise, where f_i is the i^{th} element of $\mathbf{f}$.

Next, we introduce the proposed techniques for computing the residual vectors $\Delta_1\mathbf{f}$ and $\Delta_2\mathbf{f}$ in Subsections IV-A and IV-B respectively, the convergence analysis of the proposed **AOT** framework in Subsection IV-C.

A. Label Deficiency

In this subsection, we introduce our proposed techniques to solve for residual vector $\Delta_1\mathbf{f}$, which addresses label deficiency. Based on eq. (4), this involves the minimization of both Q_1 and Q_2.

To instantiate Q_1, notice that the key to semi-supervised learning is the *consistency* assumption [11]. When we have access to small amount of labeled data and lots of unlabeled data, the classification function can be enforced to be sufficiently smooth on the intrinsic structure of the data manifold. According to the consistency assumption, if two examples are similar to each other, they should belong to the same class. So in a scenario where the examples are similar and the corresponding pseudo labels are different, the overall classification vector $\mathbf{f}$ should address the discrepancy in the class labels. More specifically, we have,

$$\begin{aligned} Q_1 &= \frac{1}{2}\sum_{i,j=1}^{m+n} \mathbf{W}_{ij}\Big(\frac{\mathbf{f}_i^0 + \Delta_1\mathbf{f}_i + \Delta_2\mathbf{f}_i}{\sqrt{\mathbf{D}_i}} - \frac{\mathbf{f}_j^0 + \Delta_1\mathbf{f}_j + \Delta_2\mathbf{f}_j}{\sqrt{\mathbf{D}_j}}\Big)^2 \\ &+ \sum_{i=1}^{m} \mu_1(\mathbf{f}_i^0 + \Delta_1\mathbf{f}_i + \Delta_2\mathbf{f}_i - \mathbf{y}_i)^2 \\ &= \frac{1}{2}(\mathbf{f}^0 + \Delta_1\mathbf{f} + \Delta_2\mathbf{f})^T(\mathbf{I} - \mathbf{S})(\mathbf{f}^0 + \Delta_1\mathbf{f} + \Delta_2\mathbf{f}) \\ &+ \mu_1||\mathbf{f}_L^0 + \Delta_1\mathbf{f}_L + \Delta_2\mathbf{f}_L - \mathbf{y}_L||^2 \end{aligned} \quad (6)$$

where $\mu_1 > 0$ is the regularization parameter. The objective function in eq. (6) has two terms. The first term is the smoothness constraint which ensures the class labels of the similar examples are similar to each other. The second term is the regularizer constraint which ensures that the optimal classification function should not change too much from the class labels of the labeled examples.

On the other hand, to instantiate Q_2, we enforce the residual vector $\Delta_1\mathbf{f}$ to be sparse. The sparsity constraint ensures that this residual vector is non-zero only when the corresponding example contains changed relationship between features and class labels, or the example has new features. To be specific, we add the elastic-net regularizer to enforce sparsity in the residual vector $\Delta_1\mathbf{f}$ as follows:

$$Q_2(\Delta_1\mathbf{f}) = \mu_2||\Delta_1\mathbf{f}||_1 + (1 - \mu_2)||\Delta_1\mathbf{f}||_2^2 \quad (7)$$

where μ_2 is the elastic-net coefficient. As the L_1 norm term in the sparse regularizer $Q_2(\Delta_1\mathbf{f})$ is not continuously differentiable and discontinuous at $\Delta_1\mathbf{f}_i = 0$, we employ the proximal gradient descent [12] to estimate the residual vector $\Delta_1\mathbf{f}$. As shown in eq. (4), the combined cost function to address the label deficiency is the sum of regularization term and the sparsity constraint given as follows:

$$Q_{\text{LabDef}}(\Delta_1\mathbf{f}, \Delta_2\hat{\mathbf{f}}) = Q_1(\Delta_1\mathbf{f}, \Delta_2\hat{\mathbf{f}}) + Q_2(\Delta_1\mathbf{f}) \quad (8)$$

where $\Delta_2\hat{\mathbf{f}}$ is the fixed distribution shift residual vector.

It can be seen that, the component Q_1 is a differentiable convex function, detailed proof is omitted due to space constraints. Also, the elastic-net sparsity constraint term Q_2 is closed, convex and non-differentiable over $\Delta_1\mathbf{f}$. The proximal gradient method can be applied to minimize the cost function in eq. (8). The proximal gradient step to compute $\Delta_1\mathbf{f}$ is $\Delta_1\mathbf{f}_k = \mathbf{prox}_{t_k Q_2}\Big(\Delta_1\mathbf{f}_{k-1} - t_k\nabla Q_2(\Delta_1\mathbf{f})\Big)$ where t_k is the step size. For the i^{th} example in the target domain, with

Algorithm 2: ADDRESSLABDEF - Addressing label deficiency

```
Input: (1) The normalized affinity matrix $\mathbf{S}$ for the $m+n$ examples; (2) The
       noisy pseudo-labels $\mathbf{f}^0$ from the off-the-shelf classifier; (3) The
       residual vector for distribution shift, $\Delta_2\mathbf{f}$; (4) The max iteration
       number $K$
Output: $\Delta_1\mathbf{f}$: The residual vector to address label deficiency.
$l_0 = 1$, $\eta = 2$
$\Delta_1\mathbf{f}_0 \leftarrow \mathbf{0}$, $\gamma_1 = \Delta_1\mathbf{f}_0$ and $t_1 = 1$
for $k \leftarrow 1$ to $K$ do
    $\hat{l} = \eta^i l_{k-1}$
    while $Q_{\text{LabDef}}(\Delta_1\mathbf{f}, \Delta_2\hat{\mathbf{f}}) > G(prox_{\lambda l}(\gamma_{k-1}), \gamma_{k-1})$ do
        $i \leftarrow i+1$
        $\hat{l} = \eta^i l_{k-1}$
    end
    $l_k = \hat{l}$; $\Delta_1\mathbf{f}_k = prox_{\lambda l}(\gamma_k)$; $t_{k+1} = \frac{1+\sqrt{1+4t_k * t_k}}{2}$
    $\gamma_{k+1} = \Delta_1\mathbf{f}_k + \frac{t_k - 1}{t_{k+1}}(\Delta_1\mathbf{f}_k - \Delta_1\mathbf{f}_{k-1})$
end
return $\Delta_1\mathbf{f}_K$
```

elastic-net coefficient μ_2, the proximal mapping for the elastic-net regularizer Q_2 is $\mathbf{prox}_{t_k Q_2}(\mathbf{f}_i) = \left(\frac{1}{\mu_2+2t-2t\mu_2}\right)(\mathbf{f}_i - t)_+ - (-\mathbf{f}_i - t)_-$. The residual vector $\Delta_1\mathbf{f}$ is iteratively computed through proximal gradient descent using a variant of fast iterative shrinkage thresholding algorithm [13].

The algorithm to address label deficiency is illustrated in Algo. 2. The algorithm takes as input the normalized affinity matrix $\mathbf{S}$ for the $m+n$ examples from the target domain, the noisy pseudo-labels $\mathbf{f}^0$ from the off-the-shelf classifier, the residual vector for distribution shift, $\Delta_2\mathbf{f}$ and the max iteration number K. It outputs the residual vector $\Delta_1\mathbf{f}$ for addressing label deficiency. In the algorithm, we first initialize the parameters, and set the initial label deficiency residual vector to $\Delta_1\mathbf{f}_0 = \mathbf{0}$, a zero vector. The Lipschitz constant for the iteration is computed through line search using the proximal gradient mapping. For any $l > 0$, consider the proximal gradient mapping at any given point γ is given by $G_l(\Delta_1\mathbf{f}, \gamma) := Q_1(\gamma) + \nabla Q_1(\gamma)^T(\Delta_1\mathbf{f} - \gamma) + \frac{l}{2}||\Delta_1\mathbf{f} - \gamma||^2 + Q_2$ where l is the Lipschitz constant. Considering $\mathbf{L} = \mathbf{I} - \mathbf{S}$, the term $G_l(\Delta_1\mathbf{f}, \gamma)$ can be computed from $Q_1(\Delta_1\mathbf{f})$ and $\nabla Q_1(\Delta_1\mathbf{f})$ terms. In each iteration the Lipschitz constant for the iteration is computed and the residual vector $\Delta_1\mathbf{f}$ is updated through proximal gradient descent steps. As shown in [13], the proposed variant of fast iterative shrinkage thresholding algorithm (Algo. 2) ensures the cost function Q_{LabDef} is monotonically decreasing and converges to the global optimal $\Delta_1\mathbf{f}^*$.

B. Distribution Shift

In traditional machine learning, often the data distribution of training and test data is considered to be the same. When the distributions are different, the trained classification model may not perform well on the test data. In source free domain adaptation, the off-the-shelf classifier is trained on a data set with a different distribution from the given data set $\mathcal{D}_L \cup \mathcal{D}_U$ in the target domain. This leads to a distribution shift as the class conditional distribution in the target domain is different from the training data used by off-the-shelf classifiers, potentially changing the optimal predicted labels. The inconsistency in the class labels can be modeled as a residual vector $\Delta_2\mathbf{f}$. Similar as in the last subsection, the cost function to address the distribution shift is the sum of the regularization term Q_1 and Q_3, which measures the prediction loss on the labeled examples from the target domain:

$$\begin{aligned} Q_{\text{DisShift}}(\Delta_1\hat{\mathbf{f}}, \Delta_2\mathbf{f}) =&\ Q_1(\Delta_1\hat{\mathbf{f}}, \Delta_2\mathbf{f}) + Q_3(\Delta_2\mathbf{f}) \\ =&\ Q_1(\Delta_1\hat{\mathbf{f}}, \Delta_2\mathbf{f}) + \frac{1}{m}\sum_i^m \Big(\mathbf{y}_i - f_i^0 - \Delta_1\hat{f}_i - \Delta_2 f_i\Big)^2 \end{aligned} \tag{9}$$

where $\Delta_1\hat{\mathbf{f}}$ is the fixed label deficiency residual vector. Notice that the cost function Q_{DisShift} is smooth and $\nabla Q_{\text{DisShift}}$ exists for $\Delta_2\mathbf{f} \in \mathbb{R}^{m+n}$. And the term $\nabla Q_{\text{DisShift}}$ can be computed as follows:

$$\nabla Q_{\text{DisShift}} = 2(\mathbf{I} - \mathbf{S})\Delta_2\mathbf{f} + 2\mu_1\mathbf{I}(\mathbf{f}^0 + \Delta_1\mathbf{f} + \Delta_2\mathbf{f} - \mathbf{y}) \tag{10}$$

We employ the gradient boosting approach to compute the residual vector $\Delta_2\mathbf{f}$ that minimizes this cost function. Like other boosting methods, gradient boosting combines a set of *weak learners* into a single strong learner in an iterative fashion. The algorithm for the gradient boosting is shown in Algo. 3. Using the labeled examples, we train a set of gradient boosted regressors and update the residual function $\Delta_2\mathbf{f}$ for all the examples $\Delta_2 f_i = \mathbf{F}(x_i), i \in 1 \ldots (m+n)$ and where $\Delta_2 f_i$ is the i^{th} element of $\Delta_2\mathbf{f}$. The gradient boosted regressor is an ensemble of SVM tree regressors trained on the m labeled examples.

Algorithm 3: ADDRESSDISSHIFT - Addressing distribution shift

```
Input: (1) Example feature matrices $\mathbf{X}_L$ and $\mathbf{X}_U$; (2) Noisy pseudo-labels
       $\mathbf{f}^0$ from the off-the-shelf classifier; (3) Residual vector for label
       deficiency, $\Delta_1\hat{\mathbf{f}}$; (4) Max iterations $K$
Output: $\Delta_2\mathbf{f}$: The residual vector to address distribution shift.
Initialize $\mathbf{F}_0$
for $k \leftarrow 1$ to $K$ do
    $\mathbf{r}_k = -\nabla Q_{\text{DisShift}}(\Delta\hat{\mathbf{f}}, \Delta_2\mathbf{f}_k)$
    Learn a base learner $\mathbf{h}_k$ on labeled examples
    $\gamma_k = \underset{\gamma}{argmin} \sum_i^m \nabla Q_{\text{DisShift}}(\Delta_1\hat{f}_i, \mathbf{F}_{k-1} + \gamma\mathbf{h}_k(x_i))$
    $\mathbf{F}_k = \mathbf{F}_{k-1} + \gamma_k\mathbf{h}_k$
end
for $i \leftarrow 1$ to $m+n$ do
    $\Delta_2 f_i = \mathbf{F}_k(x_i)$
end
return $\Delta_2\mathbf{f}$
```

Algorithm **ADDRESSDISSHIFT** (Algo. 3) shows the details for computing the residual vector $\Delta_2\mathbf{f}$ to address the distribution shift. The input to the algorithm are the example feature matrices for the labeled examples $\mathbf{X}_L$, and the unlabeled examples $\mathbf{X}_U$, the noisy pseudo-labels $\mathbf{f}^0$ from the off-the-shelf classifier, the residual vector for label deficiency, $\Delta_1\mathbf{f}$. The gradient boosting is performed by fitting all the labeled examples $\mathcal{D}_L$ to an SVM regressor and the residual value is computed for all the unlabeled examples $\mathcal{D}_U$. Finally, the algorithm outputs the residual vector $\Delta_2\mathbf{f}$.

C. Convergence of AOT

In this subsection we formally discuss the convergence of the proposed **AOT** algorithm. As discussed earlier in **AOT** algorithm (Algo. 1), we employ an alternative minimization strategy to compute the residual vectors $\Delta_1\mathbf{f}$ and $\Delta_2\mathbf{f}$. We follow the existing work [14] to prove the convergence of the proposed alternative minimization framework in Theorem. 1.

Theorem 1. *Let $\Delta_1\mathbf{f}_k, \Delta_2\mathbf{f}_k$ be the sequence generated by the proposed alternating minimization based* **AOT** *framework. Then for any $k > 0$, $L_1 > 0$, $L_2 > 0$ and for finite values of L_1 and L_2, the rate of convergence is given by*

$$Q(\Delta_1\mathbf{f}_{k+1}, \Delta_2\mathbf{f}_k) - Q(\Delta_1\mathbf{f}^*, \Delta_2\mathbf{f}^*) \leq ||G^1_{L_1}||.||\Delta_1\mathbf{f}_k + 1 - \Delta_1\mathbf{f}^*|| \tag{11}$$

$$Q(\Delta_1\mathbf{f}_k, \Delta_2\mathbf{f}_{k+1}) - Q(\Delta_1\mathbf{f}^*, \Delta_2\mathbf{f}^*) \leq ||G^2_{L_2}||.||\Delta_2\mathbf{f}_k + 1 - \Delta_2\mathbf{f}^*|| \tag{12}$$

where $G^1_{L_1}$ and $G^1_{L_1}$ is the proximal gradient mapping, $\Delta_1\mathbf{f}^*$ and $\Delta_2\mathbf{f}^*$ are the local optimal residual functions. With the above rate of convergence, the residual functions $\Delta_1\mathbf{f}_k$ and $\Delta_2\mathbf{f}_k$ computed iteratively converge to $\Delta_1\mathbf{f}^*$ and $\Delta_2\mathbf{f}^*$ respectively.

Proof. The cost function for the manifold regularization term $Q_1(\Delta_1\mathbf{f}, \Delta_2\mathbf{f})$ is a continuously differentiable convex function over domain of Q_2, $\mathbb{R}^{m+n}$ and over domain of Q_3, $\mathbb{R}^{m+n}$. The gradient of Q_1 is (uniformly) Lipschitz continuous with respect to $\Delta_1\mathbf{f}$ over the domain of Q_2 with constant $L_1 \in (0, \infty)$. Also, the gradient of Q_1 is (uniformly) Lipschitz continuous with respect to $\Delta_2\mathbf{f}$ over domain of Q_3 with constant $L_2 \in (0, \infty)$. Therefore, $||\nabla_1 Q_1(\Delta_1\mathbf{f} + \mathbf{d_1}) - \nabla_1 Q_1(\Delta_1\mathbf{f})|| \leq L_1||\mathbf{d_1}||$ and $||\nabla_2 Q_1(\Delta_2\mathbf{f} + \mathbf{d_2}) - \nabla_2 Q_1(\Delta_2\mathbf{f})|| \leq L_2||\mathbf{d_2}||$ where the Lipschitz constants $L_1 = L_2 = 2tr((1+\mu_1)\mathbf{I} - \mathbf{S}) \geq 0$, $\mathbf{d_1} \in \mathbb{R}^{m+n}$, $\mathbf{d_2} \in \mathbb{R}^{m+n}$, $\Delta_1\mathbf{f} + \mathbf{d_1}$ and $\Delta_2\mathbf{f} + \mathbf{d_2}$ is in domain of Q_2 and Q_3 respectively. The proposed alternating minimization framework **AOT** adheres to the framework proposed in the paper Beck(2015) [14]. From Lemma 3.4 in the alternating minimization framework proposed in Beck(2015) [14], the sequence $\Delta_1\mathbf{f}_k, \Delta_2\mathbf{f}_k$ generated by the proposed **AOT** framework converges to $\Delta_1\mathbf{f}^*, \Delta_2\mathbf{f}^*$. Due to space constrains, the detailed proof is omitted. □

V. Experimental Results

A. Experiment Setup

Data sets: The performance of the proposed **AOT** framework is evaluated on seven real world data sets. The statistics of all the data sets are shown in Table I. The Stanford sentiment classification tool is used to compute the off-the-shelf classification ratings for all the text data sets. The details of the data sets are as follows: (1) **IMDB movie reviews** [15]: A binary sentiment classification data set; (2) **Amazon fine food reviews** [16]: A binary sentiment classification data set, where all the reviews with 4-5 star ratings are considered as positive and reviews with 1-2 star ratings are considered negative; (3) **Cats and dogs images** [17]: A binary image classification data set. The data from Imagenet [18] with synsets *cats* and *dogs* are used to train the off-the-shelf classifier; (4) **News articles**: News articles related to illegal immigration and cartel wars in Mexico have been crawled from various news websites from United States and Mexico. The binary classification task for this data set is to identify whether the content of the news article is related to *illegal immigration* or *cartel wars*. The news articles in Spanish are translated to English using Google translation service, and used to train the black-box off-the-shelf classifier. For textual features, tf-idf feature vector as a bag of words on n-grams $n = \{1, 2, 3, 4\}$ were extracted for each review. For the images, SIFT features [19] represented as a tf-idf feature vector on Bag of Visual Words (BoVW). The BoVW are computed through K-Means on SIFT descriptors for each image. The number of clusters for cats and dogs data is set to 600. The cluster size is chosen based on the 10-fold cross validation.

Data set	Type	# of examples
Binary data sets		
IMDB movie reviews	Text	10000
Amazon fine food reviews	Text	10000
Cats and dogs images	Image	3000
News articles	Text	1395

TABLE I: Statistics of the six data sets

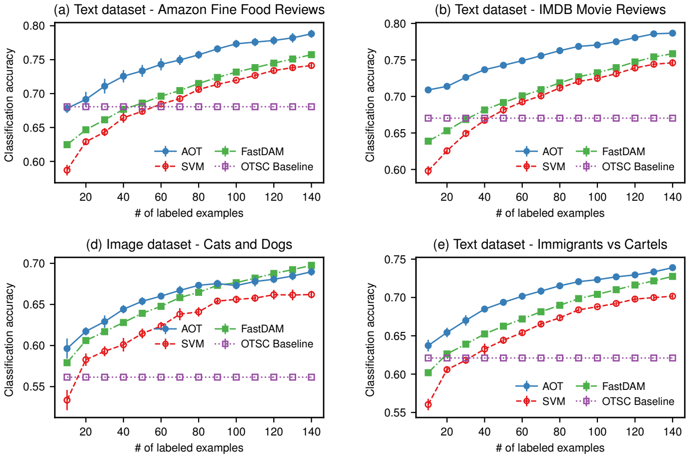

Fig. 1: Figures (a)-(d): Classification accuracy on all the binary data sets with 10-140 labeled examples

Comparison methods: The effectiveness of the proposed framework is demonstrated by comparing with the baseline off-the-shelf classifier and the strong baseline of SVM. The various methods compared in the experiments and their setup is as follows: (1) **FastDAM**: Fast Domain Adaptation Machine [20]. To compare with the proposed **AOT** framework, only one set of classification labels from the off-the-shelf classifier were considered for FastDAM; (2) **SVM**: Strong baseline SVM trained on the known labeled examples from the target domain; (3) **OTSC**: Off-the-shelf classifier. The Stanford sentiment classification toolkit is used as the off-the-shelf classifier for the binary sentiment classification data sets. For the image data sets, a logistic regression model trained on the similar images as the target domain is used as the off-the-shelf classifier.

B. *Effectiveness of AOT*

In this subsection, the effectiveness of the proposed **AOT** framework is evaluated by comparing with other methods. For all the experiments, the regularization parameters in the label deficiency **AddressLabDef** are set to $\mu_1 = 0.7$ and $\mu_2 = 0.5$. For all the experiments, the results are reported after 30 different runs on randomly sampled data from the training set. The effectiveness of the proposed **AOT** approach is evaluated from a sample of 10-140 labeled examples for the binary data sets, 20-280 examples for the multi-class data sets.

From the results in the Fig. 1, the proposed **AOT** framework performs better than all the competitors on both the text and image data sets. Its performance is very close to that of FastDAM on both image data sets. This is because the labels generated by the baseline off-the-shelf classifier are very noisy on the image data sets, and the gain achieved by the proposed **AOT** framework is limited by the quality of the labels generated by the off-the-shelf classifier.

C. *Two Stage Analysis*

We analyze the benefit of addressing label deficiency and distribution shift individually. The number of labeled examples is set to $m = 140$. We evaluate the performance of algorithms AddressLabDef, AddressDisShift and **AOT** on *Amazon fine food reviews* and *Cats and dogs* data sets. From Fig. 2, it can be observed that addressing label deficiency through manifold regularization alone is more helpful than addressing distribution shift. Also combining both algorithms performs better than the performance of the individual algorithms. This demonstrates the power of combining both algorithms together for better adaptation results.

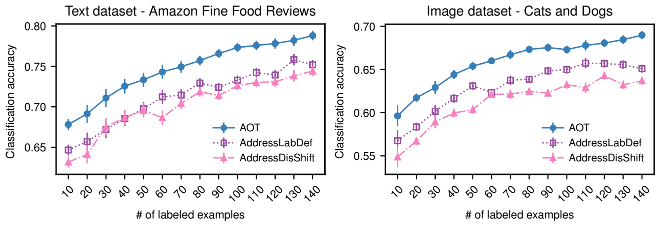

Fig. 2: Two stage analysis for the *Amazon fine foods* and *Cats and dogs* binary data sets.

D. *Sensitivity Analysis*

In this subsection, we analyze the influence of hyper-parameters on the proposed **AOT** framework. We analyze the influence of hyper-parameters through grid search. Both parameters μ_1 and μ_2 in the objective function taking values in the interval $[0, 1]$ are analyzed. In general, the performance of the proposed framework is robust to small perturbations in the parameters. Furthermore, it was observed that the parameter μ_1 which controls the influence of the regularizer on the labeled examples gives good results with higher values $\mu_1 \geq 0.6$ and performs poorly with smaller values. Also, the parameter μ_2 which controls the sparsity has a better accuracy for a balance elastic net regularizer around $\mu_2 = 0.5$.

VI. Conclusion

In this paper we propose **AOT** , a generic framework for source free domain adaptation, which aims to adapt an off-the-shelf classifier to the target domain without having access to the source domain training data. In **AOT** , we explicitly address the two main challenges, label deficiency and distribution shift by introducing two residual vectors in the optimization framework. Furthermore, we propose a variant of iterative shrinkage approach to estimate the residual vectors that converges quickly. Also, the drift in the class distribution is corrected through gradient boosting. Empirical study demonstrates the effectiveness and efficiency of our **AOT** framework on real world data sets for text and image classification.

References

[1] "Caffe model zoo." [Online]. Available: https://github.com/BVLC/caffe/wiki/Model-Zoo

[2] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP nlp toolkit," in *ACL*, 2014.

[3] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, vol. 1. IEEE, 2001.

[4] H. Wu, H. Wang, and C. Zong, "Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora," in *ACL*. ACL, 2008, pp. 993–1000.

[5] M. Long, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," 14 Feb. 2016.

[6] A. R. Nelakurthi, H. Tong, R. Maciejewski, N. Bliss, and J. He, "User-guided cross-domain sentiment classification," in *SDM*. SIAM, 2017.

[7] J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive svms," in *MM '07*, ser. MM '07. New York, NY, USA: ACM, 2007, pp. 188–197.

[8] L. Duan, I. W. Tsang, and D. Xu, "Domain transfer multiple kernel learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, Mar. 2012.

[9] Z. Lu, Y. Zhu, S. J. Pan, E. W. Xiang, Y. Wang, and Q. Yang, "Source free transfer learning for text classification," *AAAI*, 2014.

[10] B. Chidlovskii, S. Clinchant, and G. Csurka, "Domain adaptation in the absence of source domain data," in *SIGKDD*. ACM, 2016.

[11] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency." in *NIPS*, vol. 16, no. 16, 2003.

[12] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[13] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[14] A. Beck, "On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes," *SIAM Journal on Optimization*, vol. 25, no. 1, 2015.

[15] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *ACL*. Portland, Oregon, USA: ACL, June 2011.

[16] J. J. McAuley and J. Leskovec, "From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews," in *WWW*. ACM, 2013.

[17] J. Elson, J. J. Douceur, J. Howell, and J. Saul, "Asirra: A captcha that exploits interest-aligned manual image categorization," in *CCS*. ACM, October 2007.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[19] Y. Zhou, A. R. Nelakurthi, and J. He, "Unlearn what you have learned: Adaptive crowd teaching with exponentially decayed memory learners," in *SIGKDD*. ACM, 2018.

[20] L. Duan, I. W. Tsang, D. Xu, and T.-S. Chua, "Domain adaptation from multiple sources via auxiliary classifiers," in *ICML*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 289–296.