Adaptation Regularization: A General Framework for Transfer Learning

Mingsheng Long, Jianmin Wang, Guiguang Ding, Sinno Jialin Pan, and Philip S. Yu, *Fellow, IEEE*

Abstract—Domain transfer learning, which learns a target classifier using labeled data from a different distribution, has shown promising value in knowledge discovery yet still been a challenging problem. Most previous works designed adaptive classifiers by exploring two learning strategies independently: distribution adaptation and label propagation. In this paper, we propose a novel transfer learning framework, referred to as Adaptation Regularization based Transfer Learning (ARTL), to model them in a unified way based on the structural risk minimization principle and the regularization theory. Specifically, ARTL learns the adaptive classifier by simultaneously optimizing the structural risk functional, the joint distribution matching between domains, and the manifold consistency underlying marginal distribution. Based on the framework, we propose two novel methods using Regularized Least Squares (RLS) and Support Vector Machines (SVMs), respectively, and use the Representer theorem in reproducing kernel Hilbert space to derive corresponding solutions. Comprehensive experiments verify that ARTL can significantly outperform state-of-the-art learning methods on several public text and image datasets.

Index Terms—Transfer learning, adaptation regularization, distribution adaptation, manifold regularization, generalization error

1 Introduction

It is very difficult, if not impossible, to induce a supervised classifier without any labeled data. For the emerging domains where labeled data are sparse, to save the manual labeling efforts, one may expect to leverage abundant labeled data available in a related source domain for training an accurate classifier to be reused in the target domain. Recently, the literature has witnessed an increasing interest in developing *transfer learning* [1] methods for cross-domain knowledge transfer problems. Transfer learning has proven to be promising in many real-world applications, e.g., text categorization [2], [3], sentiment analysis [4], [5], image classification [6] and retrieval [7], video summarization [8], and collaborative recommendation [9].

Recall that the probability distributions in different domains may change tremendously and have very different statistical properties, e.g., mean and variance. Therefore, one major computational issue of transfer learning is how to reduce the difference in distributions between the source and target data. Recent works aim to discover a good feature representation across domains, which can simultaneously reduce the distribution difference and preserve the important properties of the original data [10]. Under the new feature representation, standard supervised learning algorithms can be trained on source domain and reused on target domain [11], [12]. Pan *et al*. [11] proposed Maximum Mean Discrepancy Embedding (MMDE), in which the MMD [13] distance measure for comparing different distributions is explicitly minimized. Si *et al*. [12] proposed a general Transfer Subspace Learning (TSL) framework, in which the Bregman divergence is imposed as a regularization to a variety of subspace learning methods, e.g., PCA and LDA. Another line of works aims to directly construct an adaptive classifier by imposing the distance measure as a regularization to supervised learning methods, e.g., SVMs [14]–[17]. However, these methods only utilized the source domain labeled data to train a classifier. We show that such labeled data can be further explored to reduce the difference in the *conditional* distributions across domains. Also, these methods only utilized the target domain unlabeled data to reduce the difference in the *marginal* distributions across domains. We show that these unlabeled data can be further explored to boost classification performance.

It is noteworthy that, in some real-world scenarios, only minimizing difference in marginal distributions between domains is not good enough for knowledge transfer, since the discriminative directions of the source and target domains may still be different [10], [18]. Therefore, another major computational issue of transfer learning is how to further explore marginal distributions to potentially match the discriminative directions between domains. In this direction, the unlabeled data may often reveal the underlying truth of the target domain [19]–[21]. Bruzzone *et al*. [19] proposed Domain Adaptation Support Vector Machine (DASVM), which extended Transductive SVM (TSVM) to

- *M. Long is with the School of Software, Tsinghua University, Beijing 100084, and with the Department of Computer Science, Tsinghua University, Beijing 100084, China. E-mail: longming-sheng@gmail.com.*
- *J. Wang, and G. Ding are with the School of Software, Tsinghua University, Beijing 100084, China. E-mail: {jimwang, dinggg}@tsinghua.edu.cn.*
- *S. J. Pan is with the Institute of Infocomm Research, Singapore 138632. E-mail: jspan@i2r.a-star.edu.sg.*
- *P. S. Yu is with the Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607 USA. E-mail: psyu@uic.edu.*

Manuscript received 15 Aug. 2012; revised 19 June 2013; accepted 20 June 2013. Date of publication 30 June 2013; date of current version 7 May 2014.
Recommended for acceptance by J. Bailey.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier 10.1109/TKDE.2013.111

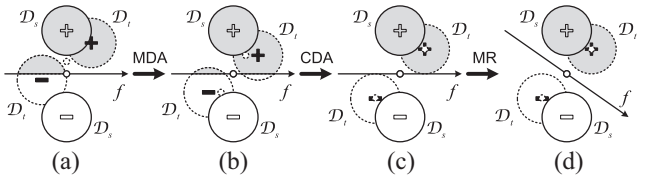


Fig. 1. Motivation of ARTL. f: hyperplane; $\mathcal{D}_S$: source domain; $\mathcal{D}_t$: target domain; ○: domain/class centroid; MDA: marginal distribution adaptation; CDA: conditional distribution adaptation; MR: manifold regularization. (a) Original domains. (b) After MDA. (c) After CDA. (d) After MR.

progressively classify the unlabeled target data and simultaneously remove some labeled source data. Bahadori *et al*. [20] proposed Latent Transductive Transfer Learning (LATTL) to combine subspace learning and transductive classification (TSVM) in a unified framework. However, all these methods adopt TSVM as building block, which is difficult to solve and is not natural for out-of-sample data [22]. In addition, these methods do not minimize the difference between the *conditional* distributions across domains.

Based on the aforementioned discussions, we summarize the computational issues of transfer learning in Fig. 1 and highlight our motivation. Given a labeled source domain $\mathcal{D}_s$ and an unlabeled target domain $\mathcal{D}_t$ as in subplot (a), we can see that hyperplane f trained on $\mathcal{D}_s$ cannot discriminate $\mathcal{D}_t$ correctly due to substantial distribution difference. Similar to most previous works, we minimize the distance between the *marginal* distributions in subplot (b), i.e., the sample moments of the two domains are drawn closer. Then hyperplane f can classify $\mathcal{D}_t$ more correctly. Noteworthily, it is indispensable to minimize the distance between the *conditional* distributions as in subplot (c), which can make the intra-class centroids close and the inter-class centroids more separable. Finally, as shown in subplot (d), it is important to maximize the manifold consistency underlying the marginal distributions, which can "rotate" hyperplane f to respect the groundtruth of the target data. This motivates us to design a general framework to integrate all these learning objectives.

In this paper, we propose a general transfer learning framework, referred to as Adaptation Regularization based Transfer Learning (ARTL), to model the joint distribution adaptation and manifold regularization in a unified way underpinned by the structural risk minimization principle and the regularization theory. More specifically, ARTL learns an adaptive classifier by simultaneously optimizing the structural risk functional, the joint distribution matching between both marginal and conditional distributions, and the manifold consistency of the marginal distribution. The contributions of this paper are summarized as follows.

- To cope with the considerable change between data distributions from different domains, ARTL aims to minimize the structural risk functional, joint adaptation of both marginal and conditional distributions, and the manifold regularization. To the best of our knowledge, ARTL is the first *semi-supervised* domain transfer learning framework which can explore all these learning criteria simultaneously. In particular, ARTL remains simple by introducing only one additional term (parameter) compared with the state-of-the-art graph-based semi-supervised learning framework [22].
- Many standard supervised methods, e.g., RLS and SVMs, can be incorporated into the ARTL framework to tackle domain transfer learning. A revised Representer theorem in the Reproducing Kernel Hilbert Space (RKHS) is presented to facilitate easy handling of optimization problems.
- Under the ARTL framework, we further propose two novel methods, i.e., ARRLS and ARSVM, respectively. Both of them are convex optimization problems enjoying the global optimal solutions.
- Comprehensive experiments on text (Reuters-21578 and 20-Newsgroups) and image (PIE, USPS, and MNIST) datasets verify the effectiveness of the ARTL framework in real-world applications.

The remainder of the paper is organized as follows. We start by reviewing related works in Section 2. In Section 3, we present the ARTL framework, the two methods ARRLS and ARSVM, and the analysis of time complexity. In Sections 4 and 5, we theoretically analyze the generalization error bound of ARTL, and conduct empirical studies on real-world datasets, respectively. Finally, we conclude the paper in Section 6.

2 Related Work

In this section, we discuss previous works on transfer learning that are most related to our work, and highlight their differences. According to literature survey [1], most previous methods can be roughly organized into two categories: *instance reweighting* [23], [24] and *feature extraction*. Our work belongs to the feature extraction category, which includes two subcategories: *transfer subspace learning* and *transfer classifier induction*.

2.1 Transfer Subspace Learning

These methods aim to extract a shared subspace in which the distributions of the source and target data are drawn close. Typical learning strategies includes:

1) *Correspondence Learning*, which first identifies the correspondence among features and then explores this correspondence for transfer subspace learning [4], [5];
2) *Property Preservation*, which extracts shared latent factors between domains by preserving the important properties of the original data, e.g., statistical property [2], [25], geometric structure [26]–[28], or both [3];
3) *Distribution Adaptation*, which learns a shared subspace where the distribution difference is explicitly reduced by minimizing predefined distance measures, e.g., MMD or Bregman divergence [10]–[12], [29].

2.2 Transfer Classifier Induction

These methods aim to directly design an adaptive classifier by incorporating the adaptation of different distributions

through model regularization. For easy discussion, the learning strategies of these methods are summarized as below. Our ARTL framework belongs to this subcategory, with substantial extensions.

1) *Subspace Learning + Classifier Induction:* These methods simultaneously extract a shared subspace and train a supervised [30] or semi-supervised classifier [20] in this subspace. The advantage is that the subspace and classifier can establish mutual reinforcement. Different from these methods, ARTL does not involve subspace learning and thus is more generic.
2) *Distribution Adaptation + Classifier Induction:* These methods directly integrate the minimization of distribution difference as a regularization term to the standard supervised classifier [15]–[17], [19]. But all these methods only minimize the distance between the *marginal* distributions. Different from these methods, ARTL minimizes the distance between both the marginal and *conditional* distributions. Our work also explores manifold structure to improve performance.
3) *Feature Replication + Co-Regularization:* In these methods, the distribution difference is firstly reduced through feature replication, then both the source and target classifiers are required to agree on the unlabeled target data [31]. These methods require some labeled data in target domain, which is not required by ARTL.
4) *Parameter Sharing + Manifold Regularization:* This strategy is explored by semi-supervised multi-task learning methods [32], [33], which aim to improve the performance of multiple related tasks by exploring the common structure through a common prior. However, these methods ignore the distribution adaptation between multiple tasks, which is different from ARTL.
5) *Kernel Matching + Manifold Regularization:* These methods simultaneously perform classifier induction, kernel matching, and manifold preservation [21]. The differences between these methods and ARTL are that: 1) these methods do not reduce the distance between *conditional* distributions; 2) kernel matching is usually formulated as an integer program, which is difficult to solve; and 3) it is difficult to encode kernel matching as a regularization to standard classifiers, as a result a Representer theorem is missing for these methods.

3 Adaptation Regularization Based Transfer Learning Framework

In this section, we first define the problem setting and learning goal for domain transfer learning. After that, we present the proposed general framework, ARTL. Based on the framework, we propose two methods using RLS and SVMs, and derive learning algorithms using the Representer theorem in RKHS. Finally, we analyze the computational complexity of the algorithms.

TABLE 1
Notations and Descriptions Used in this Paper

Notation	Description	Notation	Description
$\mathcal{D}_s, \mathcal{D}_t$	source/target domain	$\mathbf{X}$	data matrix
n, m	#examples in $\mathcal{D}_s/\mathcal{D}_t$	$\mathbf{Y}$	label matrix
d, C	#shared features/classes	$\mathbf{K}$	kernel matrix
p	#nearest neighbors	$\mathbf{w}, \boldsymbol{\alpha}$	classifier parameters
σ	shrinkage regularization	$\mathbf{E}$	label indicator matrix
λ	MMD regularization	$\mathbf{M}$	MMD matrix
γ	manifold regularization	$\mathbf{L}$	graph Laplacian matrix

3.1 Problem Definition

Notations, which are frequently used in this paper, are summarized in Table 1.

Definition 1 (Domain). *[1] A domain $\mathcal{D}$ is composed of a d-dimensional feature space $\mathcal{X}$ and a marginal probability distribution $P(\mathbf{x})$, i.e., $\mathcal{D} = \{\mathcal{X}, P(\mathbf{x})\}$, where $\mathbf{x} \in \mathcal{X}$.*

In general, if two domains $\mathcal{D}_s$ and $\mathcal{D}_t$ are different, then they may have different feature spaces or marginal distributions, i.e., $\mathcal{X}_s \neq \mathcal{X}_t \vee P_s(\mathbf{x}_s) \neq P_t(\mathbf{x}_t)$.

Definition 2 (Task). *[1] Given domain $\mathcal{D}$, a task $\mathcal{T}$ is composed of a label space $\mathcal{Y}$ and a prediction function $f(\mathbf{x})$, i.e., $\mathcal{T} = \{\mathcal{Y}, f(\mathbf{x})\}$, where $y \in \mathcal{Y}$, and $f(\mathbf{x}) = Q(y|\mathbf{x})$ can be interpreted as the conditional probability distribution.*

In general, if two tasks $\mathcal{T}_s$ and $\mathcal{T}_t$ are different, then they may have different label spaces or conditional distributions, i.e., $\mathcal{Y}_s \neq \mathcal{Y}_t \vee Q_s(y_s|\mathbf{x}_s) \neq Q_t(y_t|\mathbf{x}_t)$.

Definition 3 (Domain Transfer Learning). *Given labeled source domain $\mathcal{D}_s = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ and unlabeled target domain $\mathcal{D}_t = \{\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+m}\}$, the goal of domain transfer learning is to learn a target prediction function f_t:$\mathbf{x}_t \mapsto y_t$ with low expected error on $\mathcal{D}_t$, under the assumptions $\mathcal{X}_s = \mathcal{X}_t$, $\mathcal{Y}_s = \mathcal{Y}_t$, $P_s(\mathbf{x}_s) \neq P_t(\mathbf{x}_t)$, and $Q_s(y_s|\mathbf{x}_s) \neq Q_t(y_t|\mathbf{x}_t)$.*

To address domain transfer learning problems directly by estimating the distribution densities is challenging. Although the marginal distribution $P_t(\mathbf{x}_t)$ can be estimated using kernel density estimate (KDE) [18], it is impossible for the conditional distribution $Q_t(y_t|\mathbf{x}_t)$ since there are no labeled data in the target domain. Most previous works thus assume that there exists a proper feature transformation F such that $P_s(F(\mathbf{x}_s)) = P_t(F(\mathbf{x}_t))$, $Q_s(y_s|F(\mathbf{x}_s)) \approx Q_t(y_t|F(\mathbf{x}_t))$. The transformation F can be inferred by minimizing the distribution distance between *marginal* distributions, and preserving properties of original data [10].

In this paper, we put forward several justifications.

- It is insufficient to minimize only the distribution distance between the marginal distributions. The distribution distance between the *conditional* distributions should also be explicitly minimized.
- It is useful to further explore the marginal distributions. Noteworthily, preserving manifold consistency underlying the marginal distribution can benefit us from semi-supervised learning [22].
- It may be more generic to explore the data distributions in original feature space or kernel

space, instead of various dimension-reduced subspaces.

Based on these justifications, we propose our general ARTL framework in the following section.

3.2 General Framework

We design the general ARTL framework underpinned by the structural risk minimization principle and the regularization theory. Specifically, we aim to optimize three complementary objective functions as follows:

1) Minimizing the structural risk functional on the source domain labeled data $\mathcal{D}_s$;
2) Minimizing the distribution difference between the joint probability distributions J_s and J_t;
3) Maximizing the manifold consistency underlying the marginal distributions P_s and P_t.

Suppose the prediction function (i.e., classifier) be $f = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x})$, where $\mathbf{w}$ is the classifier parameters, and $\phi{:}\mathcal{X} \mapsto \mathcal{H}$ is the feature mapping function that projects the original feature vector to a Hilbert space $\mathcal{H}$. The learning framework of ARTL is formulated as

$$\begin{aligned} f \;=\; & \arg\min_{f \in \mathcal{H}_K} \sum_{i=1}^{n} \ell\left(f\left(\mathbf{x}_i\right), y_i\right) + \sigma \left\|f\right\|_K^2 \\ & + \lambda D_{f,K}\left(J_s, J_t\right) + \gamma M_{f,K}\left(P_s, P_t\right), \end{aligned} \tag{1}$$

where K is the kernel function induced by ϕ such that $\left\langle \phi\left(\mathbf{x}_i\right), \phi\left(\mathbf{x}_j\right)\right\rangle = K\left(\mathbf{x}_i, \mathbf{x}_j\right)$, and σ, λ, and γ are positive regularization parameters. We interpret each term of Framework (1) in the following subsections.

3.2.1 Structural Risk Minimization

Our ultimate goal is to learn an adaptive classifier for the target domain $\mathcal{D}_t$. To begin with, we can induce a standard classifier f on the labeled source domain $\mathcal{D}_s$. We adopt the structural risk minimization principle [34], and minimize the *structural risk functional* as

$$f = \arg\min_{f \in \mathcal{H}_K} \sum_{i=1}^{n} \ell\left(f\left(\mathbf{x}_i\right), y_i\right) + \sigma \left\|f\right\|_K^2, \tag{2}$$

where $\mathcal{H}_K$ is a set of classifiers in the kernel space, $\|f\|_K^2$ is the squared norm of f in $\mathcal{H}_K$, σ is the *shrinkage* regularization parameter, and ℓ is the loss function that measures the fitness of f for predicting the labels on training samples. Two widely-used loss functions are the hinge loss for SVMs $\ell = \max\left(0, 1 - y_i f\left(\mathbf{x}_i\right)\right)$, and the squared loss for RLS $\ell = \left(y_i - f\left(\mathbf{x}_i\right)\right)^2$.

3.2.2 Joint Distribution Adaptation

Unfortunately, the standard classifier f inferred by (2) may not generalize well to the target domain $\mathcal{D}_t$, since the structural risk minimization principle requires the training and test data to be sampled from identical probability distribution [34]. Thus the first major computational issue is how to minimize the distribution distance between the joint probability distributions J_s and J_t. By probability theory, $J = P \cdot Q$, thus we seek to minimize the distribution distance 1) between the marginal distributions P_s and P_t, and 2) between the conditional distributions Q_s and Q_t, simultaneously.

Marginal Distribution Adaptation: We minimize $D_{f,K}(P_s, P_t)$, the distance between marginal distributions P_s and P_t. Since directly estimating probability densities is nontrivial, we resort to explore nonparametric statistics. We adopt empirical *Maximum Mean Discrepancy* (MMD) [11], [13] as the distance measure, which compares different distributions based on the distance between the sample means of two domains in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, namely

$$\mathrm{MMD}_{\mathcal{H}}^2\left(\mathcal{D}_s, \mathcal{D}_t\right) = \left\| \frac{1}{n}\sum_{i=1}^{n} \phi\left(\mathbf{x}_i\right) - \frac{1}{m}\sum_{j=n+1}^{n+m} \phi\left(\mathbf{x}_j\right) \right\|_{\mathcal{H}}^2,$$

where $\phi{:}\mathcal{X} \mapsto \mathcal{H}$ is the feature mapping. To make MMD a proper regularization for the classifier f, we adopt the *projected* MMD [15], which is computed as

$$D_{f,K}\left(P_s, P_t\right) = \left\| \frac{1}{n}\sum_{i=1}^{n} f\left(\mathbf{x}_i\right) - \frac{1}{m}\sum_{j=n+1}^{n+m} f\left(\mathbf{x}_j\right) \right\|_{\mathcal{H}}^2, \tag{3}$$

where $f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x})$, and K is the kernel function induced by ϕ such that $\left\langle \phi\left(\mathbf{x}_i\right), \phi\left(\mathbf{x}_j\right)\right\rangle = K\left(\mathbf{x}_i, \mathbf{x}_j\right)$.

Conditional Distribution Adaptation: We minimize $D_{f,K}(Q_s, Q_t)$, the distance between conditional distributions Q_s and Q_t. Since calculating the nonparametric statistics of $Q_s(y_s|\mathbf{x}_s)$ and $Q_t(y_t|\mathbf{x}_t)$ is difficult, we resort to explore the nonparametric statistics of $Q_s(\mathbf{x}_s|y_s)$ and $Q_t(\mathbf{x}_t|y_t)$ instead, which can well approximate $Q_s(y_s|\mathbf{x}_s)$ and $Q_t(y_t|\mathbf{x}_t)$ when sample sizes are large. Unfortunately, it is impossible to calculate the sample moments of $Q_t(\mathbf{x}_t|y_t)$ w.r.t. each class (class centroids), since there are no labels in the target domain data. In this paper, we propose to use the *pseudo* target labels predicted by some supervised classifiers (e.g., SVMs) trained on the source domain labeled data. Though many of the pseudo target labels may be incorrect due to substantial distribution difference, we assume that the pseudo class centroids calculated by them may reside not far apart from the true class centroids. Therefore, we can use both true and pseudo labels to compute the projected MMD w.r.t. each class $c \in \{1, \ldots, C\}$ and make the intra-class centroids of two distributions $Q_s(\mathbf{x}_s|y_s)$ and $Q_t(\mathbf{x}_t|y_t)$ closer in $\mathcal{H}$

$$D_{f,K}^{(c)}(Q_s, Q_t) = \left\| \frac{1}{n^{(c)}} \sum_{\mathbf{x}_i \in \mathcal{D}_s^{(c)}} f(\mathbf{x}_i) - \frac{1}{m^{(c)}} \sum_{\mathbf{x}_j \in \mathcal{D}_t^{(c)}} f\left(\mathbf{x}_j\right) \right\|_{\mathcal{H}}^2, \tag{4}$$

where $\mathcal{D}_s^{(c)} = \left\{\mathbf{x}_i{:}\mathbf{x}_i \in \mathcal{D}_s \wedge y\left(\mathbf{x}_i\right) = c\right\}$ is the set of examples belonging to class c in the source data, $y\left(\mathbf{x}_i\right)$ is the true label of $\mathbf{x}_i$, and $n^{(c)} = |\mathcal{D}_s^{(c)}|$. Correspondingly, $\mathcal{D}_t^{(c)} = \left\{\mathbf{x}_j{:}\mathbf{x}_j \in \mathcal{D}_t \wedge \hat{y}\left(\mathbf{x}_j\right) = c\right\}$ is the set of examples belonging to class c in the target data, $\hat{y}\left(\mathbf{x}_j\right)$ is the pseudo (predicted) label of $\mathbf{x}_j$, and $m^{(c)} = |\mathcal{D}_t^{(c)}|$.

Integrating (3) and (4) leads to the regularization for *joint distribution adaptation*, computed as follows

$$D_{f,K}\left(J_s, J_t\right) = D_{f,K}\left(P_s, P_t\right) + \sum_{c=1}^{C} D_{f,K}^{(c)}\left(Q_s, Q_t\right). \tag{5}$$

By regularizing (2) with (5), the sample moments of both the marginal and conditional distributions are drawn closer

in $\mathcal{H}$. It is noteworthy that, if we use an adaptive classifier to obtain the pseudo labels, then we can usually obtain a more accurate labeling for the target data, which can further boost classification accuracy. ARTL can readily integrate any base classifiers by (5).

3.2.3 Manifold Regularization

In domain transfer learning, there are both labeled and unlabeled data. Since by using (5) we can only match the sample moments between different distributions, but we expect that knowledge of the marginal distributions P_s and P_t can be further exploited for better function learning. In other words, the unlabeled data may often reveal the underlying truth of the target domain, e.g., the sample variances. By the *manifold assumption* [22], if two points $\mathbf{x}_s, \mathbf{x}_t \in \mathcal{X}$ are close in the intrinsic geometry of the marginal distributions $P_s(\mathbf{x}_s)$ and $P_t(\mathbf{x}_t)$, then the conditional distributions $Q_s(y_s|\mathbf{x}_s)$ and $Q_t(y_t|\mathbf{x}_t)$ are similar. Under geodesic smoothness, the *manifold regularization* is computed as

$$\begin{aligned} M_{f,K}(P_s, P_t) &= \sum_{i,j=1}^{n+m} \left(f(\mathbf{x}_i) - f\left(\mathbf{x}_j\right)\right)^2 W_{ij} \\ &= \sum_{i,j=1}^{n+m} f(\mathbf{x}_i) L_{ij} f\left(\mathbf{x}_j\right), \end{aligned} \tag{6}$$

where $\mathbf{W}$ is the graph affinity matrix, and $\mathbf{L}$ is the normalized graph Laplacian matrix. $\mathbf{W}$ is defined as

$$W_{ij} = \begin{cases} \cos\left(\mathbf{x}_i, \mathbf{x}_j\right), & \text{if } \mathbf{x}_i \in \mathcal{N}_p\left(\mathbf{x}_j\right) \vee \mathbf{x}_j \in \mathcal{N}_p\left(\mathbf{x}_i\right) \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

where $\mathcal{N}_p(\mathbf{x}_i)$ is the set of p-nearest neighbors of point $\mathbf{x}_i$. $\mathbf{L}$ is computed as $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$, where $\mathbf{D}$ is a diagonal matrix with each item $D_{ii} = \sum_{j=1}^{n} W_{ij}$.

By regularizing (2) with (6), the marginal distributions can be fully exploited to maximize the consistency between the predictive structure of f and the intrinsic manifold structure of the data. This can substantially match the discriminative hyperplanes between domains.

3.3 Learning Algorithms

We extend standard algorithms (RLS and SVMs) under the ARTL framework with different choices of loss functions ℓ. The major difficulty lies in that the kernel mapping $\phi: \mathcal{X} \mapsto \mathcal{H}$ may have infinite dimensions. To solve (1) effectively, we need to reformulate it by using the following revised Representer theorem.

Theorem 1 (Representer Theorem). *[22], [35] The minimizer of optimization problem* (1) *admits an expansion*

$$f(\mathbf{x}) = \sum_{i=1}^{n+m} \alpha_i K(\mathbf{x}_i, \mathbf{x}) \quad \textit{and} \quad \mathbf{w} = \sum_{i=1}^{n+m} \alpha_i \phi(\mathbf{x}_i) \tag{8}$$

in terms of the cross-domain labeled and unlabeled examples, where K is a kernel induced by ϕ, α_i is a coefficient.

We focus on reformulating the regularization. By incorporating Equation (8) into Equation (5), we have

$$\begin{aligned} D_{f,K}(J_s, J_t) &= \operatorname{tr}\left(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\mathbf{M}_0\mathbf{K}\boldsymbol{\alpha}\right) + \sum_{c=1}^{C} \operatorname{tr}\left(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\mathbf{M}_c\mathbf{K}\boldsymbol{\alpha}\right) \\ &= \operatorname{tr}\left(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\mathbf{M}\mathbf{K}\boldsymbol{\alpha}\right) \quad \text{with} \quad \mathbf{M} = \sum_{c=0}^{C} \mathbf{M}_c, \end{aligned} \tag{9}$$

where $\mathbf{K} \in \mathbb{R}^{(n+m)\times(n+m)}$ is kernel matrix with $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{n+m})$ is classifier parameters. $\mathbf{M}_c, c \in \{0, 1, \ldots, C\}$ are MMD matrices computed as

$$(M_c)_{ij} = \begin{cases} \frac{1}{n^{(c)}n^{(c)}}, & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_s^{(c)} \\ \frac{1}{m^{(c)}m^{(c)}}, & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_t^{(c)} \\ \frac{-1}{n^{(c)}m^{(c)}}, & \begin{cases} \mathbf{x}_i \in \mathcal{D}_s^{(c)}, \mathbf{x}_j \in \mathcal{D}_t^{(c)} \\ \mathbf{x}_j \in \mathcal{D}_s^{(c)}, \mathbf{x}_i \in \mathcal{D}_t^{(c)} \end{cases} \\ 0, & \text{otherwise,} \end{cases} \tag{10}$$

where $n^{(c)}, m^{(c)}, \mathcal{D}_s^{(c)}, \mathcal{D}_t^{(c)}, c \in \{1, \ldots, C\}$ are defined as (4). For clarity, we can also compute $\mathbf{M}_0$ with (10) if substituting $n^{(0)} = n, m^{(0)} = m, \mathcal{D}_s^{(0)} = \mathcal{D}_s, \mathcal{D}_t^{(0)} = \mathcal{D}_t$.

Similarly, by incorporating (8) into (6), we obtain

$$M_{f,K}(P_s, P_t) = \operatorname{tr}\left(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\mathbf{L}\mathbf{K}\boldsymbol{\alpha}\right). \tag{11}$$

With (9) and (11), we can readily implement new algorithms under ARTL by extending RLS and SVMs.

3.3.1 ARRLS: ARTL Using Squared Loss

Using squared loss $\ell\left(f(\mathbf{x}_i), y_i\right) = \left(y_i - f(\mathbf{x}_i)\right)^2$, the structural risk functional can be formulated as follows

$$\begin{aligned} &\sum_{i=1}^{n} \ell\left(f(\mathbf{x}_i), y_i\right) + \sigma \|f\|_K^2 \\ &= \sum_{i=1}^{n+m} E_{ii}\left(y_i - f(\mathbf{x}_i)\right)^2 + \sigma \|f\|_K^2, \end{aligned} \tag{12}$$

where $\mathbf{E}$ is a diagonal label indicator matrix with each element $E_{ii} = 1$ if $\mathbf{x}_i \in \mathcal{D}_s$, and $E_{ii} = 0$ otherwise. By substituting Representer theorem (8) into (12), we obtain

$$\begin{aligned} &\sum_{i=1}^{n} \ell\left(f(\mathbf{x}_i), y_i\right) + \sigma \|f\|_K^2 \\ &= \left\| \left(\mathbf{Y} - \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\right)\mathbf{E} \right\|_F^2 + \sigma \operatorname{tr}\left(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\boldsymbol{\alpha}\right), \end{aligned} \tag{13}$$

where $\mathbf{Y} = [y_1, \ldots, y_{n+m}]$ is the label matrix. It is no matter that the target labels are unknown, since they are filtered out by the label indicator matrix $\mathbf{E}$. Integrating Equations (13), (9), and (11) into Framework (1), we obtain the objective for *ARRLS* based on RLS:

$$\begin{aligned} \boldsymbol{\alpha} = \underset{\boldsymbol{\alpha}\in\mathbb{R}^{n+m}}{\arg\min} &\left\| \left(\mathbf{Y} - \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\right)\mathbf{E} \right\|_F^2 \\ &+ \operatorname{tr}\left(\sigma\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\boldsymbol{\alpha} + \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}\left(\lambda\mathbf{M} + \gamma\mathbf{L}\right)\mathbf{K}\boldsymbol{\alpha}\right). \end{aligned} \tag{14}$$

Setting derivative of objective function as $\mathbf{0}$ leads to

$$\boldsymbol{\alpha} = \left(\left(\mathbf{E} + \lambda\mathbf{M} + \gamma\mathbf{L}\right)\mathbf{K} + \sigma\mathbf{I}\right)^{-1}\mathbf{E}\mathbf{Y}^{\mathrm{T}}. \tag{15}$$

Note that when $\lambda = \gamma = 0$, (15) gives zero coefficients over the joint distribution adaptation and manifold regularization and thus degenerates to standard RLS.

Multi-Class Extension: Denote $\mathbf{y} \in \mathbb{R}^C$ a label vector such that $y_c = 1$ if $y(\mathbf{x}) = c$, and $y_c = 0$ otherwise. The label matrix is $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_{n+m}] \in \mathbb{R}^{C\times(n+m)}$, and the parameter matrix is $\boldsymbol{\alpha} \in \mathbb{R}^{(n+m)\times C}$. In this way, ARRLS can be extended to multi-class problems.

3.3.2 ARSVM: ARTL Using Hinge Loss

Using hinge loss $\ell\left(f(\mathbf{x}_i), y_i\right) = \max\left(0, 1 - y_i f\left(\mathbf{x}_i\right)\right)$, the structural risk functional can be formulated as

$$\begin{aligned}&\sum_{i=1}^{n} \ell\left(f\left(\mathbf{x}_i\right), y_i\right) + \sigma \left\|f\right\|_K^2 \\ &= \sum_{i=1}^{n} \max\left(0, 1 - y_i f\left(\mathbf{x}_i\right)\right) + \sigma \left\|f\right\|_K^2. \end{aligned} \tag{16}$$

By substituting the Representer theorem (8) into (16), and integrating Equations (16), (9), and (11) into (1), we obtain the objective for *ARSVM* based on SVMs:

$$\begin{aligned}&\min_{\boldsymbol{\alpha}\in\mathbb{R}^{n+m}, \boldsymbol{\xi}\in\mathbb{R}^n} \sum_{i=1}^{n} \xi_i + \sigma \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{K} \boldsymbol{\alpha} + \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{K} \left(\lambda \mathbf{M} + \gamma \mathbf{L}\right) \mathbf{K} \boldsymbol{\alpha} \\ &\text{s.t.} \quad y_i \left(\sum_{j=1}^{n+m} \alpha_j K\left(\mathbf{x}_i, \mathbf{x}_j\right) + b \right) \geq 1 - \xi_i, i = 1, \ldots, n \\ &\qquad \xi_i \geq 0, i = 1, \ldots, n. \end{aligned} \tag{17}$$

To solve Equation (17) effectively, we follow [22] and reformulate (17) using Lagrange dual, which leads to

$$\begin{aligned}&\boldsymbol{\beta} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^n} \sum_{i=1}^{n} \beta_i - \frac{1}{2} \boldsymbol{\beta}^{\mathrm{T}} \mathbf{Q} \boldsymbol{\beta} \\ &\text{s.t.} \quad \sum_{i=1}^{n} \beta_i y_i = 0, 0 \leq \beta_i \leq \frac{1}{n}, i = 1, \ldots, n \\ &\qquad \text{with } \mathbf{Q} = \tilde{\mathbf{Y}} \tilde{\mathbf{E}} \mathbf{K} (2\sigma \mathbf{I} + 2\left(\lambda \mathbf{M} + \gamma \mathbf{L}\right) \mathbf{K})^{-1} \tilde{\mathbf{E}}^{\mathrm{T}} \tilde{\mathbf{Y}}, \end{aligned} \tag{18}$$

where $\tilde{\mathbf{Y}} = \mathrm{diag}(y_1, \ldots, y_n)$, $\tilde{\mathbf{E}} = [\mathbf{I}_n, \mathbf{0}] \in \mathbb{R}^{n\times(n+m)}$. ARSVM can be easily implemented by using a standard SVM solver with the quadratic form induced by the $\mathbf{Q}$ matrix, and then using $\boldsymbol{\beta}$ to obtain the classifier parameters by $\boldsymbol{\alpha} = (2\sigma \mathbf{I} + 2\left(\lambda \mathbf{M} + \gamma \mathbf{L}\right) \mathbf{K})^{-1} \tilde{\mathbf{E}}^{\mathrm{T}} \tilde{\mathbf{Y}} \boldsymbol{\beta}$.

The learning algorithms are summarized in Algorithm 1. To make parameters λ and γ easily tuned, we normalize graph Laplacian matrix and MMD matrix.

Algorithm 1: ARTL: Adaptation Regularization Transfer Learning Algorithms ARRLS and ARSVM

Input: Data $\mathbf{X}$, $\mathbf{Y}$; parameters p, σ, λ, γ.
Output: Adaptive classifier $f : \mathcal{X} \mapsto \mathcal{Y}$.

```
begin
    Construct MMD matrix M by Equations (9), (10), graph Laplacian L by Equation (7).
    Choose a kernel function K(x_i, x_j) and compute kernel matrix K by K_ij = K(x_i, x_j).
    Normalize M ← M / ||M||_F, L ← D^{-1/2} L D^{-1/2}.
    Compute α for ARRLS by Equation (15), for ARSVM by Equation (18) with SVM solver.
    Return adaptive classifier f by Equation (8).
```

3.4 Computational Complexity

Denote s the average number of non-zero features per example, $s \leq d$, $p \ll \min(n + m, d)$. The computational complexity of the framework consists of three parts.

1) Solving the linear systems (15) or (18) using LU decomposition requires $O((n + m)^3)$, which may be greatly reduced using the conjugate gradient method. For ARSVM, solving the SVM optimization (18) with a widely-used SVM solver [36] requires $O((n + m)^{2.3})$.
2) For constructing the graph Laplacian matrix $\mathbf{L}$, ARTL needs $O\left(s(n + m)^2\right)$, which is performed once.
3) For constructing the kernel matrix $\mathbf{K}$ and aggregate MMD matrix $\mathbf{M}$, ARTL requires $O\left(C(n + m)^2\right)$.

In summary, the computational complexity of Algorithm 1 is $O\left((n + m)^3 + (s + C)(n + m)^2\right)$ using exact computations, which is adopted in this paper. It is not difficult to speed up the algorithms using conjugate gradient methods, and this is left for our future work.

3.5 Connections to Existing Works

As discussed in Section 2, our work is substantially different from a variety of prior cross-domain learning methods such as [4]–[6], [25], [37], which do not explicitly consider distribution matching or manifold regularization. In this subsection, we will specifically distinguish our work from an insightful perspective.

Distribution Adaptation: These methods explicitly reduce distribution difference by minimizing predefined distance measures, e.g., MMD or Bregman divergence [11], [12], [14]–[17], [19]. However, they only reduce the distance between *marginal* distributions, while the distance between *conditional* distributions is not minimized. Several works considered to match both the marginal and conditional distributions [18], [29], however, they require some labeled data in the target domain, which are not required by our method. Also, the manifold structure underlying the marginal distributions is not considered in all these methods.

Manifold Regularization: These methods explicitly maximize the consistency of the induced embeddings (subspace learning) [3], [26]–[28] or classifiers (supervised learning) [22], [33] with respect to the intrinsic manifold structure. However, these methods have not explicitly reduced the distribution difference between domains and may overfit target domain data.

To our knowledge, the works most closely related to our ARTL are Graph co-regularized Transfer Learning (GTL) [3], Semi-Supervised Transfer Component Analysis (SSTCA) [10], Discriminative Feature Extraction (DFE) [30], Latent Transductive Transfer Learning (LATTL) [20], and Semi-Supervised Kernel Matching (SSKM) [21]. All these methods can be categorized as "semi-supervised transfer learning", since they have explored the combination of semi-supervised learning and transfer learning. For clear comparison, the difference between these methods is illustrated in Table 2.

TABLE 2
Comparison between Most Closely Related Works

Comparison Perspective	GTL	SSTCA	DFE	LATTL	SSKM	ARTL
Data Reconstruction	√	√		√		
Structural Risk Minimization			√	√	√	√
Marginal Adaptation		√	√		∠	√
Conditional Adaptation						√
Manifold Regularization	√	√	✓	∠	√	√
Convex Optimization						√
General Framework			√	√		√

√ unified optimization; ✓ two-step approach; ∠ alternative approach.

- GTL and SSTCA are dimensionality reduction methods where label information and manifold structure are explored only for subspace learning. Our ARTL is a framework for adaptive classifiers.
- DFE is a joint learning method for distribution adaptation and classifier training. It explores the manifold structure in a separated second step, also it does not match the conditional distributions.
- LATTL is a combination of subspace learning and transductive classification. By using Transductive SVM (TSVM) to explore both the labeled and unlabeled data, LATTL can naturally achieve a better generalization capability to the target domain. It has two weaknesses: 1) it does not explicitly reduce distribution distance; 2) its TSVM learning framework is not natural for out-of-sample data.
- SSKM is the most similar work to ours. It simultaneously considers structural risk minimization, kernel matching, and manifold preservation. The differences between SSKM and ARTL are that: 1) SSKM does not reduce the distance between conditional distributions; 2) the kernel matching is an integer programming problem and is difficult to solve; and 3) the kernel matching is not directly imposed as a regularization to the classifier, thus it does not exhibit a generic Representer theorem.

In summary, our proposed ARTL can simultaneously explore 1) structural risk minimization, 2) distribution adaptation of both the marginal and conditional distributions, and 3) manifold consistency maximization. ARTL is underpinned by the regularization theory in RKHS, and can exhibit a revised Representer theorem. Thus ARTL is a general framework in which a variety of supervised algorithms can be readily incorporated. Furthermore, ARTL is a convex optimization problem enjoying global optima. We will compare ARTL with TCA and SSKM empirically to validate its advantage.

4 Generalization Bound Analysis

We analyze the generalization error bound of ARTL on the target domain based on the structural risk on the source domain, following the approaches in [30], [38]. First, we denote the induced prediction function as $f(\mathbf{x}) = \operatorname{sgn}\left(\mathbf{w}^{\mathrm{T}}\phi\left(\mathbf{x}\right)\right)$, and the true labeling function as $h(\mathbf{x}):\mathcal{X} \mapsto \{1,-1\}$. Let $\ell(\mathbf{x})$ be a continuous loss function $\ell\left(\mathbf{x}\right) = \left|h\left(\mathbf{x}\right) - f\left(\mathbf{x}\right)\right|$, then $0 \leq \ell\left(\mathbf{x}\right) \leq 2$. First of all, the expected error of f in $\mathcal{D}_t$ is defined as

$$\epsilon_t\left(f\right) = \mathbb{E}_{\mathbf{x}\sim P_t}\left[\left|h\left(\mathbf{x}\right) - f\left(\mathbf{x}\right)\right|\right] = \mathbb{E}_{\mathbf{x}\sim P_t}\left[\ell\left(\mathbf{x}\right)\right].$$

Similarly, the expected error of f in $\mathcal{D}_s$ is defined as

$$\epsilon_s\left(f\right) = \mathbb{E}_{\mathbf{x}\sim P_s}\left[\left|h\left(\mathbf{x}\right) - f\left(\mathbf{x}\right)\right|\right] = \mathbb{E}_{\mathbf{x}\sim P_s}\left[\ell\left(\mathbf{x}\right)\right].$$

Now we present the target error bound in terms of the source risk in the following theorem, which is essentially a restatement of [38] with a slight modification.

Theorem 2. *Suppose the hypothesis space containing f is of VC-dimension d, then the expected error of f in $\mathcal{D}_t$ is bounded with probability at least $1-\delta$ by*

$$\epsilon_t\left(f\right) \leq \hat{\epsilon}_s\left(f\right) + \sqrt{\frac{4}{n}\left(d\log\frac{2en}{d} + \log\frac{4}{\delta}\right)} + D_{f,K}\left(J_s, J_t\right) + \Omega, \tag{19}$$

where e is the base of natural logarithm, $\hat{\epsilon}_s(f)$ is the empirical error of f in $\mathcal{D}_s$, and $\Omega = \inf_{f\in\mathcal{H}_K}\left[\epsilon_s\left(f\right) + \epsilon_t\left(f\right)\right]$.

From Theorem 2, the expected error in $\mathcal{D}_t$, i.e., $\epsilon_t(f)$, is bounded if we can simultaneously minimize 1) the empirical error of labeled data in $\mathcal{D}_s$, i.e., $\hat{\epsilon}_s(f)$, 2) the distribution distance between $\mathcal{D}_s$ and $\mathcal{D}_t$ in RKHS $\mathcal{H}$, i.e., $D_{f,K}(J_s, J_t)$, and 3) the adaptability of the true function h in terms of hypothesis space $\mathcal{H}_K$, i.e., Ω.

In ARTL framework, i.e., Equation (1), $\hat{\epsilon}_s(f)$ is explicitly minimized by structural risk minimization in Equation (2); $D_{f,K}(J_s, J_t)$ is explicitly minimized by distribution adaptation in Equation (5); Ω is implicitly minimized by manifold regularization in Equation (6).

Non-rigorously, we interpret why manifold regularization in Equation (6) can implicitly minimize Ω, the adaptability of the true function h in terms of the hypothesis space $\mathcal{H}_K$. First, we introduce the following theorem, which states the error bound of semi-supervised learning based on manifold regularization.

Theorem 3. *[39] Consider collection $(\mathbf{x}_i, y_i)$ for $i \in \mathcal{Z}_{n+m} = \{1,\ldots,n+m\}$. Assume that we randomly pick n distinct integers $j_1,\ldots,j_n$ from $\mathcal{Z}_{n+m}$ uniformly (without replacement), and denote it by $\mathcal{Z}_n$. Let h be the true predictor and $\hat{\mathbf{f}}(\mathcal{Z}_n)$ be the semi-supervised learner trained using labeled data in $\mathcal{Z}_n$ and unlabeled data in $\mathcal{Z}_{n+m}\backslash\mathcal{Z}_n$*

$$\hat{\mathbf{f}}\left(\mathcal{Z}_n\right) = \operatorname*{arg\,inf}_{\mathbf{f}\in\mathbb{R}^{n+m}}\left[\frac{1}{n}\sum_{i\in\mathcal{Z}_n}\ell\left(f_i, y_i\right) + \gamma\mathbf{f}^{\mathrm{T}}\mathbf{L}\mathbf{f}\right].$$

if $\left|\frac{\partial}{\partial h}\ell\left(h,y\right)\right| \leq \tau$, and $\ell(h,y)$ is convex with respect to h, then we have the generalization error bound on $\mathcal{Z}_{n+m}\backslash\mathcal{Z}_n$

$$\mathbb{E}_{\mathcal{Z}_n}\frac{1}{m}\sum_{i\in\mathcal{Z}_{n+m}\backslash\mathcal{Z}_n}\ell\left(\hat{f}_i\left(\mathcal{Z}_n\right), y_i\right) \leq \inf_{\mathbf{f}\in\mathbb{R}^{n+m}}\left[\frac{1}{n+m}\sum_{i=1}^{n+m}\ell\left(f_i, y_i\right) + \gamma\mathbf{f}^{\mathrm{T}}\mathbf{L}\mathbf{f} + \frac{\tau^2\operatorname{tr}\left(\mathbf{L}^{-1}\right)}{2\gamma n\left(n+m\right)}\right].$$

In ARTL, manifold regularization (6) is performed in RKHS $\mathcal{H}$ where the distribution distance has been minimized by Equation (5). Thus, Theorem 3 states that, the classifier $\hat{f}$ trained in a semi-supervised way on $\mathcal{D}_s \cup \mathcal{D}_t$ in RKHS $\mathcal{H}$ can be guaranteed by an error bound in $\mathcal{D}_t$. In other words, the manifold regularization (6) can implicitly minimize Ω, the adaptability of true function h in terms of the hypothesis space $\mathcal{H}_K$.

TABLE 3
Top Categories and Subcategories in 20-Newsgroups

Top Category	Subcategory	#Examples	#Features
comp	comp.graphics	970	25804
	comp.os.ms-windows.misc	963	
	comp.sys.ibm.pc.hardware	979	
	comp.sys.mac.hardware	958	
rec	rec.autos	987	
	rec.motorcycles	993	
	rec.sport.baseball	991	
	rec.sport.hokey	997	
sci	sci.crypt	989	
	sci.electronics	984	
	sci.med	987	
	sci.space	985	
talk	talk.politics.guns	909	
	talk.politics.mideast	940	
	talk.politics.misc	774	
	talk religion misc	627	

5 Experiments

In this section, we perform extensive experiments on two real-world applications (i.e., text classification and image recognition) to evaluate ARTL. Datasets and codes will be available online upon publication.

5.1 Data Preparation

5.1.1 Text Datasets

The 219 cross-domain text datasets are generated from 20-Newsgroups and Reuters-21578, which are two benchmark text corpora widely used for evaluating transfer learning algorithms [2], [10], [17], [25], [26].

20-Newsgroups[1] has approximately 20,000 documents distributed evenly in 20 different subcategories. The corpus contains four top categories *comp*, *rec*, *sci* and *talk*. Each top category has four subcategories, which are listed in Table 3. In the experiments, we can construct 6 dataset groups for binary classification by randomly selecting two top categories (one for positive and the other one for negative) from the four top categories. The 6 dataset groups are *comp vs rec*, *comp vs sci*, *comp vs talk*, *rec vs sci*, *rec vs talk*, and *sci vs talk*. Similar to the approach in [25], we set up one dataset (including source domain and target domain) for cross-domain classification as follows. For each pair of top categories P and Q (e.g., P for positive and Q for negative), their four sub-categories are denoted by P_1, P_2, P_3, P_4 and Q_1, Q_2, Q_3, Q_4, respectively. We randomly select (without replacement) two subcategories from P (e.g., P_1 and P_2) and two subcategories from Q (e.g., Q_1 and Q_2) to form a source domain, then the remaining subcategories in P and Q (i.e., P_3, P_4 and Q_3, Q_4) are selected to form a target domain. This dataset construction strategy ensures that the domains of labeled and unlabeled data are related, since they are under the same top categories. Besides, the domains are also ensured to be different, since they are drawn from different subcategories. In this way, for each dataset group P *vs* Q, we can generate $C_4^2 \cdot C_4^2 = 36$ datasets. Clearly, for each example in the generated dataset group, its class label is either P or Q. In total, we can generate 6 dataset groups consisting of $6 \cdot 36 = 216$ datasets. For fair comparison, the 216 datasets are constructed using a preprocessed version of 20-Newsgroups [2], which contains 25,804 features and 15,033 documents, with each document weighted by *term frequency-inverse document frequency* (TF-IDF).

Reuters-21578[2] has three top categories *orgs*, *people*, and *place*. Using the same strategy, we can construct 3 cross-domain text datasets *orgs vs people*, *orgs vs place* and *people vs place*. For fair comparison, we use the preprocessed version of Reuters-21578 studied in [40].

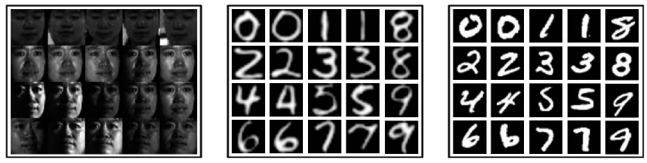


Fig. 2. Benchmark image datasets PIE, USPS, MNIST.

5.1.2 Image Datasets

USPS, MNIST and PIE (refer to Fig. 2 and Table 4) are three handwritten digits/face datasets broadly adopted in compute vision and pattern recognition.

USPS[3] dataset composes of 7,291 training images and 2,007 test images of size 16×16.

MNIST[4] dataset has a training set of 60,000 examples and a test set of 10,000 examples of size 28×28.

From Fig. 2, we see that USPS and MNIST follow different distributions. They share 10 semantic classes, with each corresponding to one digit. We construct one dataset *USPS vs MNIST* by randomly sampling 1,800 images in USPS to form the source domain, and sampling 2,000 images in MNIST to form the target domain. Then we switch the source/target pair to get another dataset *MNIST vs USPS*. We uniformly rescale all images to size 16×16, and represent each image by a 256-dimensional vector encoding the gray-scale values of all pixels. In this way, the source and target domain are ensured to share the same feature space.

PIE[5], standing for "Pose, Illumination, Expression", is a benchmark face database. It has 68 individuals with 41,368 face images sized 32×32. The images were captured by 13 synchronized cameras and 21 flashes, under varying poses, illuminations, and expressions.

In our experiments, we simply adopt the preprocessed versions of PIE[6], i.e., **PIE1** [41] and **PIE2** [42], which are generated by randomly sampling the face images from the near-frontal poses (C27) under different lighting and illumination conditions. We construct one dataset *PIE1 vs PIE2* by selecting all 2,856 images in PIE1 to form the source domain, and all 3,329 images in PIE2 to form the target domain. We switch source/target pair to get another dataset *PIE2 vs PIE1*. Thus the source and target domains are guaranteed to follow different distributions in the same feature space, due to variations in lighting and illumination.

1. http://people.csail.mit.edu/jrennie/20newsgroups
2. http://www.daviddlewis.com/resources/testcollections/reuters21578
3. http://www-i6.informatik.rwth-aachen.de/~keysers/usps.html
4. http://yann.lecun.com/exdb/mnist
5. http://vasc.ri.cmu.edu/idb/html/face
6. http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html

TABLE 4
Statistics of the 4 Benchmark Image Datasets

Dataset	Type	#Examples	#Features	#Classes
USPS	Digit	1,800	256	10
MNIST	Digit	2,000	256	10
PIE1	Face	2,856	1,024	68
PIE2	Face	3,329	1,024	68

5.2 Experimental Setup

5.2.1 Baseline Methods

We compare ARTL approaches, i.e., ARSVM and ARRLS, with eight state-of-the-art supervised and transfer learning methods for text and image classification:

- Logistic Regression (LR)
- Support Vector Machine (SVM)
- Laplacian SVM (LapSVM) [22]
- Cross-Domain Spectral Classification (CDSC) [26]
- Spectral Feature Alignment (SFA) [5]
- Transfer Component Analysis (TCA) [10]
- Large Margin Transductive TL (LMTTL) [15]
- Semi-Supervised Kernel Matching (SSKM) [21]

Specifically, LMTTL is a special case of ARTL with $\gamma = 0, C = 0$, while SSKM can be viewed as a special case of ARTL with $C = 0$. SSKM adopts a kernel matching strategy which needs an additional mapping matrix to match different kernels. Different from SSKM, ARTL seamlessly integrates the distribution adaptation term into the classifier based on the regularization theory. Note that we do not compare with [30] because their work cannot cope with thousands of training samples.

5.2.2 Implementation Details

Following [1], [10], [21], LR and SVM are trained on the labeled source data, and tested on the unlabeled target data; CDSC, SFA, and TCA are run on all data as dimensionality reduction step, then an LR classifier is trained on the labeled source data to classify the unlabeled target data; LapSVM, LMTTL, SSKM, ARSVM, and ARRLS are trained on all data in a transductive way to directly induce domain-adaptive classifiers.

Under our experimental setup, it is impossible to automatically tune the optimal parameters for the target classifier using cross validation, since we have no labeled data in the target domain. Therefore, we evaluate the eight baseline methods on our datasets by empirically searching the parameter space for the optimal parameter settings, and report the best results of each method. For LR[7] and SVM[8], we set the trade-off parameter C (i.e., $1/2\sigma$ in ARTL) by searching $C \in \{0.1, 0.5, 1, 5, 10, 50, 100\}$. For LapSVM[9], we set regularization parameters γ_A and γ_I (i.e., σ and γ in ARTL) by searching $\gamma_A, \gamma_I \in \{0.01, 0.05, 0.1, 0.5, 1, 5, 10\}$. For transfer subspace learning methods CDSC, SFA, and TCA, we set the optimal subspace dimension k by searching $k \in \{4, 8, 16, 32, 64, 128\}$. For transfer classifier induction methods LMTTL and SSKM, we set the trade-off parameter λ between the structural risk functional and the distribution adaptation term by searching $\lambda \in \{0.01, 0.1, 1, 10, 100\}$. We use linear kernel, i.e., $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$, for all kernel methods.

ARTL approaches involve four tunable parameters: shrinkage/MMD/manifold regularization parameters σ, λ, γ, and #nearest neighbors p. Sensitivity analysis validates that ARTL can achieve stable performance under a wide range of parameter values, especially for σ, λ, and p. In the comparative study, we fix $\sigma = 0.1$, $\lambda = 10$, $p = 10$, and set 1) $\gamma = 10$ for the text datasets, and 2) $\gamma = 1$ for the image datasets. In practice, we can simplify model selection by sequentially choosing optimal parameter values from the most stable ones to the most sensitive ones. Firstly, since the adaptation regularization can largely control model complexity, ARTL is very robust to σ, and we can simply choose small σ such that ARTL does not degenerate. Secondly, since distribution adaptation is inevitable for transfer learning, we choose λ such that ARTL can sufficiently match both the marginal and conditional distributions across domains. Finally, we can choose γ by following the graph-based semi-supervised learning framework [22], where p is often predetermined as KNN methods.

We use the classification *Accuracy* on the test data (unlabeled target data) as the evaluation metric, since it is widely adopted in the literature [5], [10], [17], [30]

$$Accuracy = \frac{\left|\mathbf{x}{:}\mathbf{x} \in \mathcal{D}_t \wedge f(\mathbf{x}) = y(\mathbf{x})\right|}{\left|\mathbf{x}{:}\mathbf{x} \in \mathcal{D}_t\right|},$$

where $y(\mathbf{x})$ is the groundtruth label of $\mathbf{x}$ while $f(\mathbf{x})$ is the label predicted by the classification algorithm.

5.3 Experimental Results

In this section, we compare our ARTL with the eight baseline methods in terms of classification accuracy.

5.3.1 Results of Text Classification

As 20-Newsgroups and Reuters-21578 are different in hierarchical structure, we report the results separately.

20-Newsgroups: The average classification accuracy of ARTL approaches, including ARSVM and ARRLS, and the eight baseline methods on the 6 cross-domain dataset groups (216 datasets) are illustrated in Table 5. All the detailed results of the 6 dataset groups are listed in Fig. 3, subplots (a)~(f). Each of these six figures contains the results on the 36 datasets in the corresponding group. The 36 datasets are sorted by an increasing order of the classification accuracy obtained by Logistic Regression (LR). Therefore, the x-axis in each figure can essentially indicate the degree of difficulty in cross-domain knowledge transfer. From these figures, we can make the following observations.

ARTL approaches achieve much better performance than the eight baseline methods with statistical significance. The average classification accuracy of ARRLS on the 216 datasets is **93.40%**. The performance improvement is **5.37%** compared to the best baseline method SSKM, which means a very significant error reduction of **44.86%**. Since these results are obtained from a large number of datasets, it can convincingly verify that ARTL can build robust adaptive classifiers for classifying cross-domain documents accurately.

7. http://www.csie.ntu.edu.tw/~cjlin/liblinear
8. http://www.csie.ntu.edu.tw/~cjlin/libsvm
9. http://vikas.sindhwani.org/manifoldregularization.html

TABLE 5
Average Classification Accuracy (%) on the 6 Cross-Domain Text Dataset Groups Comprising of 216 Datasets

Dataset Group	Standard Learning			Transfer Subspace Learning			Transfer Classifier Induction			
	LR	SVM	LapSVM	CDSC	SFA	TCA	LMTTL	SSKM	ARSVM	ARRLS
comp vs rec	88.37	87.51	81.93	87.95	89.73	95.12	92.15	96.06	95.10	**96.64**
comp vs sci	77.87	75.38	68.96	75.72	78.07	77.32	77.58	84.15	84.53	**86.71**
comp vs talk	96.31	95.44	95.40	97.33	95.85	97.20	94.93	97.40	97.53	**98.03**
rec vs sci	75.28	73.82	74.21	77.53	79.25	82.31	78.24	85.71	87.19	**91.02**
rec vs talk	82.28	83.27	87.44	82.14	86.98	86.58	84.55	90.15	95.99	**96.82**
sci vs talk	76.99	76.85	80.22	80.97	79.27	79.30	74.80	74.74	89.03	**91.11**
Average	82.85	82.05	81.36	83.62	84.86	86.31	83.71	88.03	91.56	**93.40**

Secondly, we observe that all the transfer learning methods can achieve better classification accuracy than the standard learning methods. A major limitation of existing standard learning methods is that they treat the data from different domains as if they were drawn from a homogenous distribution. In reality, the identical-distribution assumption does not hold in the cross-domain learning problems, and thus results in their unsatisfactory performance. It is important to notice that, the state-of-the-art semi-supervised learning method LapSVM cannot perform better than LR and SVM. Although LapSVM can explore the target data in a transductive way, it does not minimize the distribution difference between domains. Therefore, it may overfit the target data when the discriminative directions are significantly different between domains.

Thirdly, we notice that ARTL significantly outperforms CDSC, SFA, and TCA, which are state-of-the-art transfer subspace learning methods based on feature transformation. A major limitation of existing transfer subspace learning methods is that they are prone to overfitting, due to their incapability to simultaneously reduce the difference in both marginal and conditional distributions between domains. Although SFA works particularly well for sentiment classification, it works fairly for text classification, and the reason is that SFA only explores feature co-occurrence for feature alignment without considering feature frequency, which is effective for low-frequency sentiment data but not effective for high-frequency text data. ARTL addresses these limitations and can achieve much better results.

Fourthly, we observe that ARTL achieves much better performance than LMTTL and SSKM. Notice that, LMTTL and SSKM are typical transfer classifier induction methods, which can induce a supervised/semi-supervised classifier and meanwhile minimize the distribution difference between domains. However, since the difference between the *conditional* distributions is not minimized, while the regularization terms are not imposed to the classifier, it is likely that these methods cannot fully reduce the distribution difference and may get stuck in poor local optima. ARTL achieves superior performance by alleviating these limitations.

Lastly, ARTL approaches often perform more robustly on difficult-to-classify datasets than all baseline methods. This can be observed from Figs. 3∼3, where the improvements of ARTL over the baseline methods are more remarkable on datasets in which LR performs with extremely low accuracy (below 70%).

Reuters-21578: The classification accuracy of ARTL and the baseline methods on the 3 datasets generated from Reuters-21578 are illustrated in Fig. 4(a). We observe that ARTL has outperformed, or achieved comparable performance than the baseline methods.

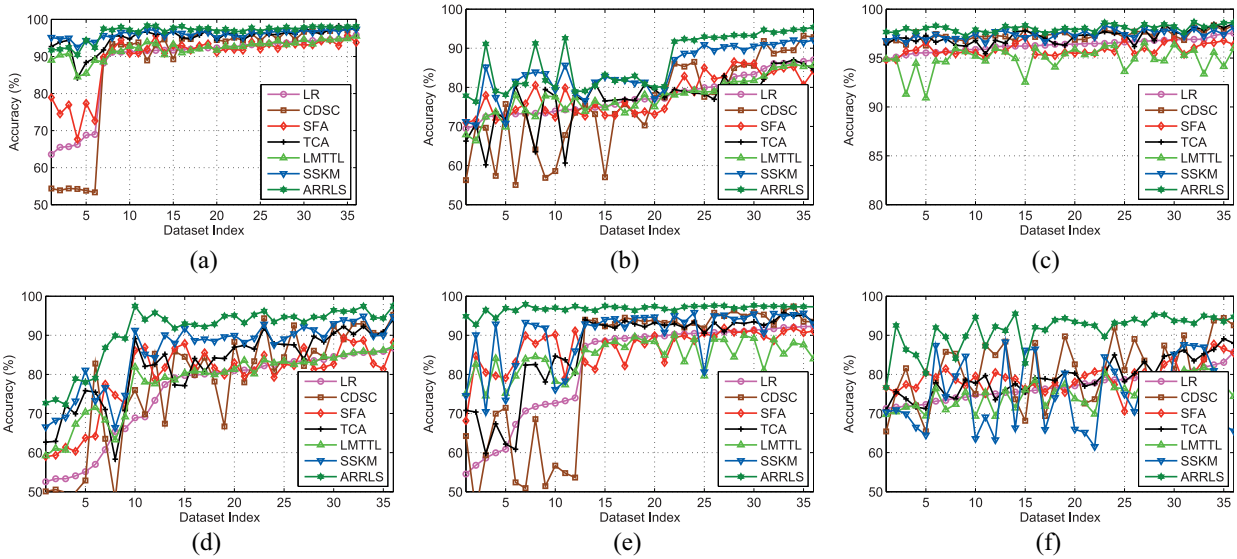
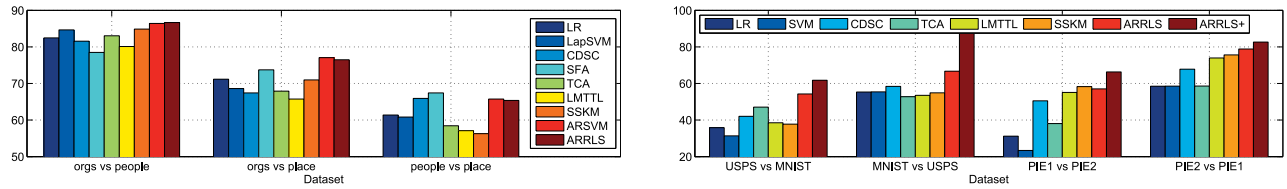


Fig. 3. Classification accuracy of LR, CDSC, SFA, TCA, LMTTL, SSKM, and ARRLS on the 216 text datasets: (a) *comp vs rec*. (b) *comp vs sci*. (c) *comp vs talk*. (d) *rec vs sci*. (e) *rec vs talk*. (f) *sci vs talk*.



(a) (b)

Fig. 4. Classification accuracy of LR, SVM, LapSVM, CDSC, TCA, LMTTL, SSKM, ARSVM, ARRLS, ARRLS+: (a) Reuters-21578. (b) Image datasets.

We notice that, Reuters-21578 is more challenging than 20-Newsgroups, since each of its top categories consists of many subcategories, i.e., clusters or subclasses. Therefore, it is more difficult to minimize the distribution difference by only matching the sample moments between domains. This reason can explain the unsatisfactory performance obtained by distribution adaptation methods, i.e., TCA, LMTTL, and SSKM.

By minimizing the distribution difference between both marginal and conditional distributions, ARTL can naturally match more statistical properties, i.e., both domain centroids and class centroids. Also, by maximizing the manifold consistency, ARTL can fully explore the marginal distributions, which can implicitly "rotate" the decision hyperplane to better respect the target data. In this way, ARTL can perform better on difficult datasets with many classes or subclasses.

5.3.2 Results of Image Recognition

The average classification accuracy of ARTL and the six baseline methods on the four image datasets is illustrated in Fig. 4(b). SFA is not compared since it cannot handle non-sparse image data, while LapSVM and ARSVM are not compared since their original implementations cannot deal with multi-class problems.

We notice that, the transfer subspace learning methods, e.g., CDSC, generally outperform standard LR and SVM. This is an expected result, since subspace learning methods, e.g., PCA, are very effective for image representation. Unfortunately, TCA has generally underperformed CDSC at this time. The main reasons are two-folds: 1) the MMD distance measure is not very suitable for image data, as exemplified by [12]; 2) the distribution difference is significantly large in the image datasets, resulting in the overfitting issues.

We also notice that, the transfer classifier induction methods, i.e., LMTTL and SSKM, outperform CDSC in the face datasets but underperform CDSC in the handwritten digits datasets. We conjecture the reasons as follows: 1) for the face datasets, there are 68 classes, thus transfer classifier induction methods which directly inject the labeled information into the learning procedure, are more effective; 2) for the handwritten digits datasets, data reconstruction may be a more important process to reduce the distribution difference.

In conclusion, ARTL generally outperforms all baseline methods. Therefore, we can often achieve a robust adaptive classifier, by minimizing the difference between both marginal and conditional distributions, and meanwhile preserving the manifold consistency.

5.4 Effectiveness Verification

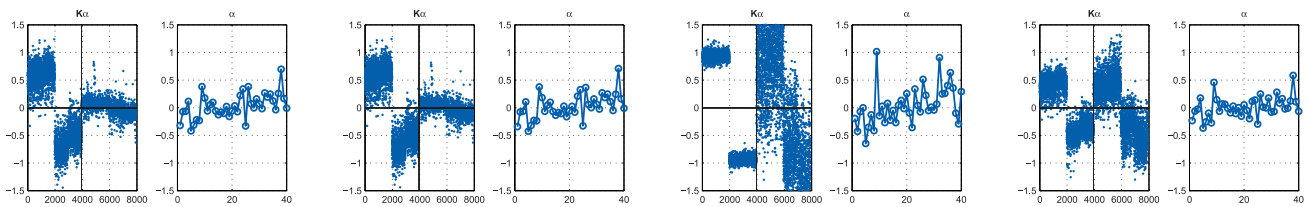
We verify effectiveness of ARTL by inspecting the impacts of base classifier and adaptation regularization.

5.4.1 Base Classifier Integration

ARTL utilizes some base classifier, e.g., SVM, to obtain the pseudo labels for the target data, through which the difference between the conditional distributions is minimized. Unsurprisingly, if we use some adaptive classifier, e.g., ARRLS, to obtain more accurate pseudo labels for the target data, then we can match the conditional distributions more accurately and further boost the classification accuracy. It is very interesting that ARTL can accept its outputs as inputs to iteratively improve itself. We denote this alternatingly enhanced version of ARRLS as ARRLS+. We run ARRLS+ on the image datasets, and show its classification accuracy in Fig. 4(b). Similar results on other datasets are omitted due to space limitation. We note that ARRLS+ has significantly outperformed ARRLS by **10.60%**, which verifies ARTL can naturally integrate base classifiers.

5.4.2 Adaptation Regularization

To inspect the effectiveness of each criterion, we run ARRLS on a randomly selected dataset, e.g., *rec vs sci 1*, by removing one term from its objective function.



(a) (b) (c) (d)

Fig. 5. Classification predictions $\mathbf{K}\boldsymbol{\alpha}$ and classifier parameters $\boldsymbol{\alpha}$ output by ARRLS on the *rec vs sci 1* dataset: (a) $C = 0$. (b) $\lambda = 0$. (c) $\gamma = 0$. (d) Optimal parameters.

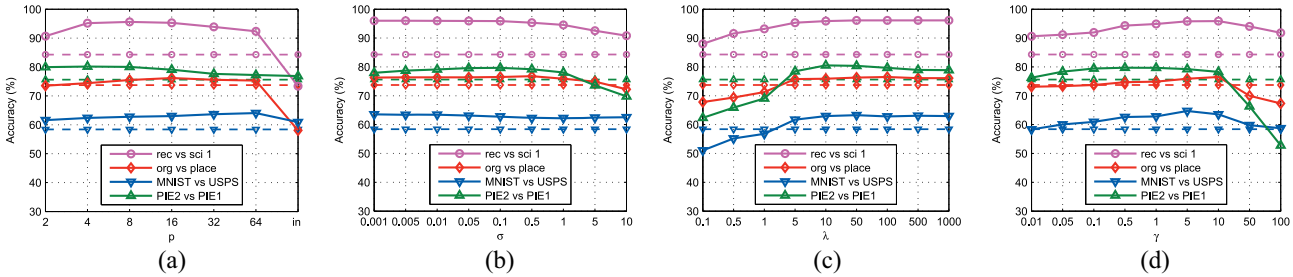


Fig. 6. Parameter sensitivity study for ARTL on selected datasets (dashed lines show the best baseline results): (a) Nearest neighbors p. (b) Shrinkage regularization σ. (c) MMD regularization λ. (d) Manifold regularization γ.

First, we remove the conditional distribution adaptation term by setting $C = 0$ as in Fig. 5(a). In this case, we cannot even find a clear decision hyperplane for the target data, i.e., the target data are not well separated at all. This verifies the crucial role that the conditional distribution adaptation has played. Similar results can be observed from Fig. 5(b), in which we remove the whole distribution adaptation term by setting $\lambda = 0$. The similar results between $C = 0$ and $\lambda = 0$ implies that minimizing the difference between the conditional distributions is much more important than that of the marginal distributions. With conditional distribution adaptation, we can make the intra-class centroids close and the inter-class centroids more separable, as can be clearly observed from Fig. 5(d).

Secondly, we remove the manifold regularization term by setting $\gamma = 0$ as in Fig. 5(c). In this case, the predictions are scattering in a wider range than the groundtruth $[-1, 1]$. In other words, the manifold consistency underlying the target data is violated, regardless that the target data are better separated due to the distribution adaptation of both the marginal and conditional distributions. Therefore, to induce a good adaptive classifier using the ARTL framework, it is very important to preserve the manifold consistency. The importance of the manifold regularization can be observed by comparing Fig. 5(c) with Fig. 5(d).

5.5 Parameter Sensitivity

We conduct empirical parameter sensitivity analysis, which validates that ARTL can achieve optimal performance under wide range of parameter values. Due to space limitation, we randomly select one generated dataset from 20-Newsgroups, Reuters-21578, USPS & MNIST, and PIE respectively, and discuss the results.

#Nearest Neighbors p: We run ARTL with varying values of p. Theoretically, p should be neither too large nor too small, since an extremely dense graph ($p \to \infty$) will connect two examples which are not similar at all, while an extremely sparse graph ($p \to 0$) will capture limited similarity information between examples. We plot the classification accuracy w.r.t. different values of p in Fig. 6(a), which indicates a wide range $p \in [4, 64]$ for optimal parameter values.

Shrinkage Regularization σ: We run ARTL with varying values of σ. Theoretically, σ controls model complexity of the adaptive classifier. When $\sigma \to 0$, the classifier degenerates and overfitting occurs. On the contrary, when $\sigma \to \infty$, ARTL is dominated by the shrinkage regularization without fitting the input data. We plot the classification accuracy w.r.t. different values of σ in Fig. 6(b), and choose $\sigma \in [0.001, 1]$.

MMD Regularization λ: We run ARTL with varying values of λ. Theoretically, larger values of λ make distribution adaptation more effective. When $\lambda \to 0$, distribution difference is not reduced and overfitting occurs. We plot classification accuracy w.r.t. different values of λ in Fig. 6(c), and can choose $\lambda \in [5, 1000]$.

Manifold Regularization γ: We run ARTL with varying values of γ. Theoretically, larger values of γ make manifold consistency more important in ARTL. When $\gamma \to \infty$, only manifold consistency is preserved while labeled information is discarded, which is unsupervised. We plot classification accuracy w.r.t. different values of γ in Fig. 6(d), and choose $\gamma \in [0.1, 10]$.

5.6 Time Complexity

We empirically check the time complexity of all algorithms by running them on the *comp vs rec 1* dataset with 25,800 features and 8,000 documents, and show the results in Table 6. We see that ARRLS can achieve comparable time complexity as the baseline methods.

TABLE 6
Time Complexity of ARTL and the Baseline Methods

Method	Running Time (s)	Method	Running Time (s)
LR	0.05	SVM	6.79
LapSVM	44.20	CDSC	25.37
SFA	20.82	TCA	1794.90
LMTTL	604.69	SSKM	191.32
ARSVM	730.09	ARRLS	48.94

6 Conclusion

In this paper, we proposed a general framework, referred to as Adaptation Regularization based Transfer Learning (ARTL), to address cross-domain learning problems. ARTL aims to simultaneously optimize the structural risk functional, joint distribution adaptation of both the marginal and conditional distributions, and the manifold consistency. An important advantage of ARTL is that it can explore as many necessary learning objectives as possible, yet still remain simple to implement practically. Furthermore, many existing supervised learning algorithms, e.g., RLS and SVM, can be readily incorporated into the ARTL framework. ARTL is robust to the distribution

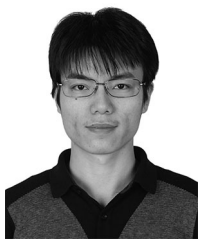
difference between domains, and can significantly improve cross-domain text/image classification problems. Extensive experiments on 219 text datasets and 4 image datasets validate that the proposed approach can achieve superior performance than state-of-the-art adaptation methods.

Acknowledgments

This work was supported in part by National HGJ Key Project (2010ZX01042-002-002), and in part by National High-Tech Development Program (2012AA040911), National Basic Research Program (2009CB320700), and National Natural Science Foundation of China (61073005, 61271394). P. S. Yu was supported in part by US NSF through Grants OISE-1129076, CNS-1115234, DBI-0960443, and in part by the US Department of Army through Grant W911NF-12-1-0066.

References

[1] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[2] F. Zhuang *et al.*, "Mining distinction and commonality across multiple domains using generative model for text classification," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 11, pp. 2025–2039, Nov. 2011.
[3] M. Long, J. Wang, G. Ding, D. Shen, and Q. Yang, "Transfer learning with graph co-regularization," in *Proc. 26th AAAI*, 2012.
[4] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *Proc. 2006 Conf. EMNLP*, 2006.
[5] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, "Cross-domain sentiment classification via spectral feature alignment," in *Proc. 19th Int. Conf. WWW*, Raleigh, NC, USA, 2010.
[6] Y. Zhu *et al.*, "Heterogeneous transfer learning for image classification," in *Proc. 25th AAAI*, 2011.
[7] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele, "What helps where—And why? Semantic relatedness for knowledge transfer," in *Proc. 23rd IEEE Conf. CVPR*, 2010.
[8] L. Li, K. Zhou, G.-R. Xue, H. Zha, and Y. Yu, "Video summarization via transferrable structured learning," in *Proc. Int. Conf. WWW*, Hyderabad, India, 2011.
[9] B. Li, Q. Yang, and X. Xue, "Transfer learning for collaborative filtering via a rating-matrix generative model," in *Proc. 26th ICML*, Montreal, Canada, 2009.
[10] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.
[11] S. J. Pan, J. T. Kwok, and Q. Yang, "Transfer learning via dimensionality reduction," in *Proc. 22nd AAAI*, 2008.
[12] S. Si, D. Tao, and B. Geng, "Bregman divergence-based regularization for transfer subspace learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 7, pp. 929–942, Jul. 2010.
[13] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Scholkopf, and A. J. Smola, "A kernel method for the two-sample problem," in *Proc. NIPS*, 2006.
[14] J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive svms," in *Proc. 15th ACM MM*, Augsburg, Germany, 2007.
[15] B. Quanz and J. Huan, "Large margin transductive transfer learning," in *Proc. 18th ACM CIKM*, Hong Kong, China, 2009.
[16] J. Tao, F.-L. Chung, and S. Wang, "On minimum distribution discrepancy support vector machine for domain adaptation," *Pattern Recognit.*, vol. 45, no. 11, pp. 3962–3984, 2012.
[17] L. Duan, I. W. Tsang, and D. Xu, "Domain transfer multiple kernel learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 465–479, Mar. 2012.
[18] E. Zhong *et al.*, "Cross domain distribution adaptation via kernel mapping," in *Proc. 15th ACM SIGKDD Int. Conf. KDD*, Paris, France, 2009.
[19] L. Bruzzone and M. Marconcini, "Domain adaptation problems: A DASVM classification technique and a circular validation strategy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 770–787, May 2010.
[20] M. T. Bahadori, Y. Liu, and D. Zhang, "Learning with minimum supervision: A general framework for transductive transfer learning," in *Proc. 11th IEEE ICDM*, 2011.
[21] M. Xiao and Y. Guo, "Semi-supervised kernel matching for domain adaptation," in *Proc. 26th AAAI*, 2012.
[22] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, Nov. 2006.
[23] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proc. 24th ICML*, Corvallis, OR, USA, 2007.
[24] J. Jiang and C. Zhai, "Instance weighting for domain adaptation in nlp," in *Proc. 45th ACL*, Prague, Czech Republic, 2007.
[25] W. Dai, G.-R. Xue, Q. Yang, and Y. Yu, "Co-clustering based classification for out-of-domain documents," in *Proc. 13th ACM SIGKDD Int. Conf. KDD*, San Jose, CA, USA, 2007.
[26] X. Ling, W. Dai, G.-R. Xue, Q. Yang, and Y. Yu, "Spectral domain-transfer learning," in *Proc. 14th ACM SIGKDD Int. Conf. KDD*, Las Vegas, NV, USA, 2008.
[27] C. Wang and S. Mahadevan, "Heterogeneous domain adaptation using manifold alignment," in *Proc. 25th AAAI*, 2011.
[28] X. Shi, Q. Liu, W. Fan, and P. S. Yu, "Transfer across completely different feature spaces via spectral embedding," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 4, pp. 906–918, Apr. 2013.
[29] B. Quanz, J. Huan, and M. Mishra, "Knowledge transfer with low-quality data: A feature extraction issue," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 10, pp. 1789–1802, Oct. 2012.
[30] B. Chen, W. Lam, I. Tsang, and T.-L. Wong, "Extracting discriminative concepts for domain adaptation in text mining," in *Proc. 15th ACM SIGKDD Int. Conf. KDD*, 2009.
[31] H. Daumé, III, A. Kumar, and A. Saha, "Co-regularization based semi-supervised domain adaptation," in *Proc. Adv. NIPS*, 2010.
[32] A. Argyriou and T. Evgeniou, "Multi-task feature learning," in *Proc. NIPS*, 2006.
[33] Q. Liu, X. Liao, and L. Carin, "Semi-supervised multitask learning," in *Proc. Adv. NIPS*, 2007.
[34] V. Vapnik, *Statistical Learning Theory*. Wiley, 1998.
[35] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Proc. 14th Annu. Conf. COLT*, Amsterdam, The Netherlands, 2001.
[36] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, 2011, [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm
[37] M. Long, J. Wang, G. Ding, W. Cheng, X. Zhang, and W. Wang, "Dual transfer learning," in *Proc. 12th SIAM Int. Conf. SDM*, 2012.
[38] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Proc. Adv. NIPS*, 2006.
[39] R. Johnson and T. Zhang, "Graph-based semi-supervised learning and spectral kernel design," *IEEE Trans. Inf. Theory*, vol. 54 , no. 1, pp. 275–288, Jan. 2008.
[40] J. Gao, W. Fan, J. Jiang, and J. Han, "Knowledge transfer via multiple model local structure mapping," in *Proc. 14th ACM SIGKDD Int. Conf. KDD*, 2008.
[41] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011.
[42] D. Cai, X. He, and J. Han, "Spectral regression: A unified approach for sparse subspace learning," in *Proc. IEEE ICDM*, Omaha, NE, USA, 2007.



Mingsheng Long received the B.S. degree in 2008 from the Department of Electrical Engineering, Tsinghua University, Beijing, China. He is a Ph.D. candidate in the Department of Computer Science and Technology, Tsinghua University. His current research interests include transfer learning, deep learning, sparse learning, and large-scale data mining.

Jianmin Wang graduated from Peking University, Beijing, China, in 1990, and received the M.E. and the Ph.D. degrees in computer software from Tsinghua University, Beijing, China, in 1992 and 1995, respectively. He is now a Professor in the School of Software, Tsinghua University. His current research interests include unstructured data management, workflow and BPM technology, benchmark for database system, information system security, and large-scale data analytics. He has published over 100 DBLP indexed papers in major journals (*TKDE, DMKD, DKE, WWWJ*, etc.) and conferences (SIGMOD, VLDB, ICDE, CVPR, AAAI, etc). He led to develop a product data/lifecycle management system, which has been implemented in hundreds of enterprises in China. He leads to develop an unstructured data management system named LaUDMS.

Guiguang Ding received the Ph.D. degree in electronic engineering from the University of Xidian, Xi'an, China. He is an Associate Professor in the School of Software, Tsinghua University, Beijing, China. His current research interests include the area of multimedia information retrieval and mining, with specific focus on visual object recognition, automatic semantic annotation, image coding and representation, and social media recommendation. He has published over 40 research papers in international conferences and journals and applied for 18 Patent Rights in China.

Sinno Jialin Pan received the Ph.D. degree in computer science from the Hong Kong University of Science and Technology (HKUST), Hong Kong. He is currently a Head (acting) of the Text Analytics Lab of the Data Analytics Department at the Institute for Infocomm Research (I^2R), Singapore. He also holds an adjunct position of Assistant Professor, Department of Computer Science at the National University of Singapore (NUS), Singapore. His current research interests include transfer learning, active learning, semi-supervised learning and their applications in text mining, pervasive computing, medical engineering, and software engineering.

Philip S. Yu received the Ph.D. degree in E.E. from Stanford University, Stanford, CA, USA. He is a Distinguished Professor in Computer Science at the University of Illinois at Chicago, Chicago, IL, USA, and holds the Wexler Chair in Information Technology. Dr. Yu is a Fellow of the ACM and the IEEE. He is the Editor-in-Chief of *ACM Transactions on Knowledge Discovery from Data*. He was the Editor-in-Chief of *IEEE Transactions on Knowledge and Data Engineering* (2001–2004). He received a Research Contributions Award from IEEE International Conference on Data Mining (2003) and a Technical Achievement Award from IEEE Computer Society (2013).