Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Fredholm Multiple Kernel Learning for Semi-Supervised Domain Adaptation

Wei Wang,[1] **Hao Wang,**[1,2] **Chen Zhang,**[1] **Yang Gao**[1]
1. Science and Technology on Integrated Information System Laboratory
2. State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing 100190, China
wangwei2014@iscas.ac.cn

Abstract

As a fundamental constituent of machine learning, domain adaptation generalizes a learning model from a *source* domain to a different (but related) *target* domain. In this paper, we focus on semi-supervised domain adaptation and explicitly extend the applied range of unlabeled target samples into the combination of distribution alignment and adaptive classifier learning. Specifically, our extension formulates the following aspects in a single optimization: 1) learning a cross-domain predictive model by developing the Fredholm integral based kernel prediction framework; 2) reducing the distribution difference between two domains; 3) exploring multiple kernels to induce an optimal learning space. Correspondingly, such an extension is distinguished with allowing for noise resiliency, facilitating knowledge transfer and analyzing diverse data characteristics. It is emphasized that we prove the differentiability of our formulation and present an effective optimization procedure based on the reduced gradient, guaranteeing rapid convergence. Comprehensive empirical studies verify the effectiveness of the proposed method.

Introduction

Conventional supervised learning aims at generalizing a model inferred from labeled training samples to test samples. For well generalization capability, it is required to collect and label plenty of training samples following the same distribution of test samples, however, which is extremely expensive in practical applications. To relieve the contradiction between generalization performance and label cost, domain adaptation (Pan and Yang 2010) has been proposed to transfer knowledge from a relevant but different *source* domain with sufficient labeled data to the *target* domain. It gains increased importance in many applied areas of machine learning (Daumé III 2007; Pan et al. 2011), including natural language processing, computer vision and WiFi localization.

Supervised domain adaptation takes advantage of both abundant labeled samples from source domain and a relatively small number of labeled samples from target domain. A line of recent work in this supervised setting (Aytar and Zisserman 2011; Daumé III and Marcu 2006; Liao, Xue, and Carin 2005; Hoffman et al. 2014) learns cross-domain classifiers by adapting traditional models, e.g., support vector machine, logistic regression and statistic classifiers, to the target domain. However, they do not explicitly explore the distribution inconsistency between the two domains, which essentially limits their capability of knowledge transfer. To this end, many semi-supervised domain adaptation methods are proposed within the classifier adaptation framework (Chen, Weinberger, and Blitzer 2011; Duan, Tsang, and Xu 2012; Sun et al. 2011; Wang, Huang, and Schneider 2014). Compared with supervised ones, they take into account unlabeled target samples to cope with the inconsistency of data distributions, showing improved classification and generalization capability. Nevertheless, most work in this area only incorporates unlabeled target samples in distribution alignment, but not in adaptive classifier learning. Large quantities of unlabeled target samples potentially contain rich information and incorporating them is desirable for noise resiliency. Moreover, some work in this semi-supervised setting is customized for certain classifiers, and the extension to other predictive models remains unclear.

In conventional supervised learning, kernel methods provide a powerful and unified prediction framework for building non-linear predictive models (Rifkin, Yeo, and Poggio 2003; Shawe-Taylor and Cristianini 2004; Yen et al. 2014). To further incorporate unlabeled samples with labeled ones in model learning, the kernel prediction framework is developed into semi-supervised setting by formulating a regularized Fredholm integral equation (Que, Belkin, and Wan 2014). Although this development is proven theoretically and empirically to be effective in noise suppression, the performance heavily depends on the choice of a single predefined kernel function. The reason is that its solution is based on Representer Theorem (Scholkopf and Smola 2001) in the Reproducing Kernel Hilbert Space (RKHS) induced by the kernel. More importantly, the kernel methods are not developed for domain adaptation, and the distribution difference will invalidate the predictive models across two domains.

In this paper, we focus on semi-supervised domain adaptation and extend the applied range of unlabeled target samples from the distribution alignment into the whole learning process. This extension further explores the structure information in target domain, enhancing robustness in complex knowledge propagation. The key challenge is to accomplish this extension through a single optimization combining the following three aspects: 1) learning a predictive model

across two domains based on Fredholm integrals utilizing (labeled) source data and (labeled and unlabeled) target data; 2) reducing the distribution difference between domains; 3) exploring a convex combination of multiple kernels to optimally induce the RKHS. In dealing with the above difficulties, the proposed algorithm named Transfer Fredholm Multiple Kernel Learning (TFMKL) has a three-fold contribution. First, compared with traditional semi-supervised domain adaptation, TFMKL learns a predictive model for noise resiliency and improved adaptation power. It is also of high expansibility due to the compatibility with many classifiers. Second, from the view of kernel prediction, the challenge of reducing distribution difference is suitably addressed. Moreover, multiple kernels are optimally combined in TFMKL, allowing for analyzing the useful data characteristics from various aspects and enhancing the interpretability of predictive model. Third, instead of employing alternate optimization technique, we prove the differentiability of our formulation and propose a simple but efficient procedure to perform reduced gradient descent, guaranteeing rapid convergence. Experimental results on a synthetic example and two real-world applications verify the effectiveness of TFMKL.

Relative Work

Domain Adaptation

In supervised domain adaptation, cross-domain classifiers (Aytar and Zisserman 2011; Hoffman et al. 2014) are learnt by using labeled source samples and a small number of labeled target samples. Meanwhile, some semi-supervised methods (Chen, Weinberger, and Blitzer 2011; Duan, Tsang, and Xu 2012; Sun et al. 2011; Wang, Huang, and Schneider 2014) are proposed by combining the transfer of classifiers with the match of distributions. This setting encourages the domain-invariant characteristics, leading to improved adaptation results for classification. Specifically, (Chen, Weinberger, and Blitzer 2011) minimizes conditional distribution difference and adapts logistic regression to target data simultaneously. Inspired from Maximum Mean Discrepancy (Borgwardt et al. 2006), the methods in (Sun et al. 2011; Wang, Huang, and Schneider 2014) match conditional or marginal distributions by transforming or re-weighting, and then use all processed labeled data to train traditional classifiers. The above methods ignore unlabeled target data in the process of learning adaptive classifiers, while the unlabeled data is desirable for robustness and noise resiliency. In (Duan, Tsang, and Xu 2012), although unlabeled target data is used during learning adaptive Support Vector Regression (SVR), it lacks theoretical support for noise suppression and is specific to SVR.

Kernel Prediction Framework

In conventional supervised setting, the kernel prediction framework (Shawe-Taylor and Cristianini 2004) specifies a positive definite kernel function $\mathbf{K}$ and then estimates a predictive function $f : \mathcal{X} \to \mathcal{Y}$ from the labeled training set $D = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\} \subset \mathcal{X} \times \mathcal{Y}$. Let $\mathcal{H}$ be the RKHS induced by $\mathbf{K}$, and this estimation is generally modeled as the following optimization problem over $\mathcal{H}$:

$$f = arg \min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} L(f(\mathbf{x}_i), y_i) + \beta \|f\|_{\mathcal{H}}^2, \tag{1}$$

where β is a tradeoff parameter, $L(z, y)$ is a risk function, e.g., *square loss*: $(z - y)^2$ and *hinge loss*: $max(1 - zy, 0)$.

In semi-supervised setting, suppose the labeled and the unlabeled training data are drawn from the distribution $P(\mathbf{X}, \mathbf{Y})$ and the marginal distribution $P(\mathbf{X})$ respectively. Associated with an outside kernel $k^P(\mathbf{x}, \mathbf{z})$, the regularized Fredholm integral $\mathcal{K}_P$ for $f(\mathbf{x})$ (Que, Belkin, and Wan 2014) is introduced to explore unlabeled data: $\mathcal{K}_P f(\mathbf{x}) = \int k^P(\mathbf{x}, \mathbf{z}) f(\mathbf{z}) P(\mathbf{z}) d\mathbf{z}$. In this term, Eq. (1) can be rewritten as: $f = arg \min \frac{1}{n} \sum_{i=1}^{n} L(\mathcal{K}_P f(\mathbf{x}_i), y_i) + \beta \|f\|_{\mathcal{H}}^2$. It has been shown that the performance of supervised algorithms could be degraded under the "noise assumption". By contrast, the f based on the Fredholm integral will be resilient to noise and closer to the optimum, because this integral provides a good approximation to the "true" data space. However, this framework is effective only when labeled and unlabeled data follow the same distribution. In addition, it relies on the single k^P and the choice of k^P heavily influences the performance.

Method

In this section, the problem in TFMKL is firstly defined. Second, we improve the kernel prediction framework and extend it to semi-supervised domain adaptation. Third, the distribution difference is minimized, which is a crucial element to gain more support from source to target domain. Fourth, the above two goals are unified together in TFMKL. Furthermore, Multiple Kernel Learning (MKL) (Bach, Lanckriet, and Jordan 2004) is exploited within the framework to improve the flexibility. Finally, we propose a reduced gradient descent procedure to update the kernel function and the predictive model simultaneously, explicitly enabling rapid convergence.

Notations and Settings

Suppose the data originates from two domains, i.e., source and target domains. The source data is a set of n_s fully labeled points $D_s = \{(\mathbf{x}_1^s, y_1^s), ..., (\mathbf{x}_{n_s}^s, y_{n_s}^s)\} \subset \mathcal{R}^d \times \mathcal{Y}$ drawn from the distribution $P_s(\mathbf{X}, \mathbf{Y})$. The target data is divided into n_l ($n_l \ll n_s$) labeled points $D_t^l = \{(\mathbf{x}_1^t, y_1^t), ..., (\mathbf{x}_{n_l}^t, y_{n_l}^t)\} \subset \mathcal{R}^d \times \mathcal{Y}$ from the distribution $P_t(\mathbf{X}, \mathbf{Y})$ and n_u ($n_u \gg n_l$) unlabeled points $D_t^u = \{\mathbf{x}_{n_l+1}^t, ..., \mathbf{x}_{n_l+n_u}^t\} \subset \mathcal{R}^d$ from the marginal distribution $P_t(\mathbf{X})$. The label set $\mathcal{Y}$ is assumed to be either $\{+1, -1\}$ in binary classification or the real line $\mathcal{R}$ in regression for simplicity, but it can be easily generalized to multi-class setting. By making full use of all available data (i.e., labeled and unlabeled points) in every learning aspect, the task of TFMKL is to find an optimal predictive model $f : \mathcal{R}^d \to \mathcal{Y}$ having low prediction error with respect to the target domain.

Adaptive Kernel Prediction Framework

Fredholm Integral on Source Domain Suppose an outside kernel $k^s(\mathbf{x}^s, \mathbf{z}^s)$ is given for source domain, where $\mathbf{x}^s$

and $\mathbf{z}^s$ follow the marginal distribution $P_s(\mathbf{X})$. As stated in the relative work, we define the Fredholm integral $\mathcal{K}_{P_s}$ as: $\mathcal{K}_{P_s} f(\mathbf{x}^s) = \int k^s(\mathbf{x}^s, \mathbf{z}^s) f(\mathbf{z}^s) P_s(\mathbf{z}^s) d\mathbf{z}^s$. Following the law of large number, $\mathcal{K}_{P_s} f(\mathbf{x}^s)$ can be approximated with the labeled data points $\mathbf{x}_1^s, ..., \mathbf{x}_{n_s}^s$, as:

$$\mathcal{K}_{P_s} f(\mathbf{x}^s) = \frac{1}{n_s} \sum_{i=1}^{n_s} k^s(\mathbf{x}^s, \mathbf{x}_i^s) f(\mathbf{x}_i^s). \tag{2}$$

Fredholm Integral on Target Domain $k^t(\mathbf{x}^t, \mathbf{z}^t)$ is given for target domain, where $\mathbf{x}^t$ and $\mathbf{z}^t$ follow $P_t(\mathbf{X})$. $\mathcal{K}_{P_t}$ is defined as: $\mathcal{K}_{P_t} f(\mathbf{x}^t) = \int k^t(\mathbf{x}^t, \mathbf{z}^t) f(\mathbf{z}^t) P_t(\mathbf{z}^t) d\mathbf{z}^t$. Associated with the labeled and the unlabeled data points $\mathbf{x}_1^t, ..., \mathbf{x}_{n_l+n_u}^t$, $\mathcal{K}_{P_t} f(\mathbf{x}^t)$ can be approximated as:

$$\mathcal{K}_{P_t} f(\mathbf{x}^t) = \frac{1}{n_l + n_u} \sum_{i=1}^{n_l+n_u} k^t(\mathbf{x}^t, \mathbf{x}_i^t) f(\mathbf{x}_i^t). \tag{3}$$

Minimizing Prediction Error To eliminate the assumption that both source and target data follow the identical distribution in traditional kernel prediction, we have constructed two Fredholm integrals on the two domains respectively. Furthermore, we adjust the relative importance of each domain by parameters λ_s and λ_t, and then estimate f by minimizing the combined prediction error:

$$\begin{aligned} f =& arg \min_{f \in \mathcal{H}} \ell(f, \mathbf{K}, D_s, D_t) = arg \min_{f \in \mathcal{H}} \{\beta \|f\|_{\mathcal{H}}^2 \\ &+ \frac{\lambda_s}{n_s} \sum_{i=1}^{n_s} L(\mathcal{K}_{P_s} f(\mathbf{x}_i^s), y_i^s) + \frac{\lambda_t}{n_l} \sum_{i=1}^{n_l} L(\mathcal{K}_{P_t} f(\mathbf{x}_i^t), y_i^t)\}, \end{aligned} \tag{4}$$

where $\mathcal{H}$ is some RKHS defined by a kernel $\mathbf{K}$. Eq. (4) intuitively considers the gap between the two domains, hence naturally boosts the performance on target domain.

Reducing Mismatch of Data Distributions

In this section, we explicitly minimize the distribution difference between domains. Following (Pan et al. 2011), the empirical Maximum Mean Discrepancy (MMD) (Borgwardt et al. 2006) is adopted as the measure of comparing marginal distributions. Specifically, given the source data D_s and the target data D_t, the distance between the data distributions of two domains in the RKHS $\mathcal{H}$ induced by the kernel $\mathbf{K}$ can be estimated as the distance between the empirical data means:

$$\begin{aligned} d_{\mathbf{K}}(D_s, D_t) &= \| \frac{1}{n_s} \sum_{i=1}^{n_s} \phi(\mathbf{x}_i^s) - \frac{1}{n_u + n_l} \sum_{i=1}^{n_u+n_l} \phi(\mathbf{x}_i^t) \|_{\mathcal{H}}^2 \\ &= tr(\Phi^T \Phi \mathbf{S}) = tr(\mathbf{K}\mathbf{S}), \end{aligned} \tag{5}$$

where $\phi : \mathcal{R}^d \to \mathcal{H}$ with $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$, $\Phi = \{\phi(\mathbf{x}_1^s), ..., \phi(\mathbf{x}_{n_s}^s), \phi(\mathbf{x}_1^t), ..., \phi(\mathbf{x}_{n_l+n_u}^t)\}$, $\mathbf{S}(i,j) = 1/n_s^2$ if $\mathbf{x}_i, \mathbf{x}_j \in D_s$; $\mathbf{S}(i,j) = 1/(n_l+n_u)^2$ if $\mathbf{x}_i, \mathbf{x}_j \in D_t$; $\mathbf{S}(i,j) = -1/n_s(n_l+n_u)$ otherwise. Based on the optimal $\mathbf{K}$ obtained by minimizing Eq. (5), the two distributions are drawn close in the induced $\mathcal{H}$.

Transfer Fredholm Multiple Kernel Learning

Semi-Supervised Learning Framework For achieving the noise resiliency and assisting the information transfer, TFMKL formulates the predictive model f and the kernel $\mathbf{K}$ in a unified framework. Taking advantage of unlabeled samples, the optimal f and $\mathbf{K}$ are found by simultaneously matching the distributions and minimizing the prediction errors on both source and target data. Based on Eq. (4) and Eq. (5), the learning framework is concisely written as:

$$[f, \mathbf{K}] = arg \min_{f, \mathbf{K}} \ell(f, \mathbf{K}, D_s, D_t) + \theta \Omega(d_{\mathbf{K}}(D_s, D_t)), \tag{6}$$

where $\Omega(\cdot)$ is a monotonic increasing function and θ is the tradeoff parameter to balance the mismatch and the errors.

Multiple Kernel Learning Instead of predefining the kernel $\mathbf{K}$ in Eq. (6) as the standard kernel prediction framework, our framework explicitly explores the optimal $\mathbf{K}$ for domain adaptation. However, directly learning a nonparametric kernel matrix by solving an semidefinite programming (Boyd and Vandenberghe 2004) is computationally prohibitive with $O(n^{6.5})$ complexity. As an efficient solution to this dilemma, MKL considers the learnt kernel as a convex combination of given (base) kernels $\mathbf{K}_m$: $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^M d_m \mathbf{K}_m(\mathbf{x}_i, \mathbf{x}_j)$, where $d_m \geq 0$ and $\sum_{m=1}^M d_m = 1$. Consequently, the problem of kernel learning is translated to the choice of optimal weights d_m, and thus the $\Omega(d_{\mathbf{K}}(D_s, D_t))$ in Eq. (6) can be rewritten into:

$$\Omega(d_{\mathbf{K}}(D_s, D_t)) = (tr((\sum_{m=1}^M d_m \mathbf{K}_m)\mathbf{S}))^2 = \mathbf{d}^T \mathbf{p}\mathbf{p}^T \mathbf{d}, \tag{7}$$

where $\mathbf{d} = [d_1, ..., d_M]^T$, $\mathbf{p} = [p_1, ..., p_M]^T$with $p_m = tr(\mathbf{K}_m \mathbf{S})$. Due to the advantages in exploring prior knowledge, describing data characteristics and enhancing interpretability, the adoption of MKL provides improved performance and generalization for our proposed algorithm.

Discussion: In contrast to previous semi-supervised domain adaptation methods, the proposed framework provides a natural way of incorporating unlabeled target data in two parts: 1) learning the adaptive predictive model f; 2) minimizing the distribution difference. The first part benefits robustness and noise resilience for domain adaptation. Different from "cluster assumption" (Chapelle, Weston, and Schölkopf 2003) or "manifold assumption" (Belkin, Niyogi, and Sindhwani 2006), the Fredholm integral is discussed under the "noise assumption" (Que, Belkin, and Wan 2014): in the neighbor of every sample, the directions with low variance are uninformative with respect to class labels and can be regarded as noise. Based on this assumption, the Fredholm integral is proven to have noise-suppression power. Specifically, when samples are polluted by noise, it will provide a more accurate estimate of the true data, and the resulting f will be more closer to the true optimum. The second part facilitates the knowledge transfer from D_s to D_t, thus f trained in $\mathcal{H}$ will make high-confidence predictions on D_t^u. Note that our domain adaptation setting is different from multi-task learning which tries to learn both target and source tasks (Evgeniou, Micchelli, and Pontil 2005).

Implementation and Optimization

TFMKL Using Square-Loss For a concrete implementation, *square loss* is applied in Eq. (6), but it is emphasized

that other loss functions can also be employed, e.g., *hinge loss* and *logistic loss*. Substituting Eq. (7) into Eq. (6), the *square loss* based formulation is:

$$\begin{aligned}\min_{f,\mathbf{d}} \ & \{\beta\|f\|_{\mathcal{H}}^2 + \frac{\lambda_s}{n_s}\sum_{i=1}^{n_s}(\mathcal{K}_{P_s}f(\mathbf{x}_i^s) - y_i^s)^2 \\ & + \frac{\lambda_t}{n_l}\sum_{i=1}^{n_l}(\mathcal{K}_{P_t}f(\mathbf{x}_i^t) - y_i^t)^2 + \theta\mathbf{d}^T\mathbf{p}\mathbf{p}^T\mathbf{d}\} \qquad (8) \\ s.t. \ & d_m \geq 0, \sum_{m=1}^{M} d_m = 1.\end{aligned}$$

One standard approach for optimizing TFMKL is to alternatively update f and $\mathbf{d}$, however, it lacks the convergence guarantees and may lead to numerical problems. Therefore, we make great efforts to propose a solution to this challenge. Generally, the cost function in Eq. (8) can be rewritten as:

$$\min_{\mathbf{d}} J(\mathbf{d}) \quad \text{with } d_m \geq 0, \sum_{m=1}^{M} d_m = 1, \qquad (9)$$

$$\begin{aligned}\text{where } J(\mathbf{d}) = \min_{f\in\mathcal{H}}\{ & \beta\|f\|_{\mathcal{H}}^2 + \frac{\lambda_s}{n_s}\sum_{i=1}^{n_s}(\mathcal{K}_{P_s}f(\mathbf{x}_i^s) - y_i^s)^2 \\ & + \frac{\lambda_t}{n_l}\sum_{i=1}^{n_l}(\mathcal{K}_{P_t}f(\mathbf{x}_i^t) - y_i^t)^2 + \theta\mathbf{d}^T\mathbf{p}\mathbf{p}^T\mathbf{d}\}.\end{aligned} \qquad (10)$$

In the following paragraphs, we will firstly prove the differentiability of $J(\cdot)$, which is at the core of the optimization. Afterwards, Problem (9) can be optimized by a reduced gradient method, ensuring the convergence to local minimum. It should be noticed that the above procedure holds for TFMKL with other loss functions, as long as the loss functions provide differentiable $J(\cdot)$. For instance, the procedure is applicable to TFMKL with *hinge loss*, because we can prove that it has a SVM-like formulation and provides a differentiable $J(\cdot)$ in a similar way as the proof of TFMKL using *square loss*.

The Differentiability of $J(\cdot)$ Although Eq. (10) is an optimization problem in a potentially infinite dimensional space $\mathcal{H}$, the following proposition reduces Eq. (10) to a finite dimensional problem.

Proposition 1. *The solution of Eq. (10) has the form:*

$$f^*(\mathbf{x}) = \sum_{i=1}^{n_s+n_l+n_u}\sum_{m=1}^{M} d_m\mathbf{K}_m(\mathbf{x},\mathbf{x}_i)v_i, \qquad (11)$$

for some $\mathbf{v} \in \mathcal{R}^{n_s+n_l+n_u}$.

The proof is similar to that of Representer Theorem (Scholkopf and Smola 2001). Given $\mathcal{K}_{P_s}$ in Eq. (2) and $\mathcal{K}_{P_t}$ in Eq. (3), Eq. (10) is translated to the quadratic optimization with the following closed-form solution $\mathbf{v}^*$ based on Proposition 1:

$$\mathbf{v}^* = (\sum_{m=1}^{M} d_m\bar{\mathbf{K}}^T\tilde{\mathbf{K}}\mathbf{K}_m + \beta\mathbf{I})^{-1}\tilde{\mathbf{K}}^T\mathbf{y}, \qquad (12)$$

where $\mathbf{y} = [y_1^s, ..., y_{n_s}^s, y_1^t, ..., y_{n_l}^t]^T$, $\bar{\mathbf{K}} = \begin{pmatrix} \frac{\mathbf{K}^s}{n_s} & \mathbf{0} \\ \mathbf{0} & \frac{\mathbf{K}^t}{n_l+n_u} \end{pmatrix}$, $\tilde{\mathbf{K}} = \begin{pmatrix} \frac{\lambda_s\mathbf{K}^s}{n_s^2} & \mathbf{0} \\ \mathbf{0} & \frac{\lambda_t\mathbf{K}^t}{(n_l+n_u)n_l} \end{pmatrix}$. Note that $(\mathbf{K}^s)_{i,j} = k^s(\mathbf{x}_i^s, \mathbf{x}_j^s)$ for $1 \leq i \leq n_s$, $1 \leq j \leq n_s$, and $(\mathbf{K}^t)_{i,j} = k^t(\mathbf{x}_i^t, \mathbf{x}_j^t)$ for $1 \leq i \leq n_l$, $1 \leq j \leq n_l + n_u$.

Algorithm 1 Transfer Fredholm Multiple Kernel Learning

Initialization:
Set $\{d_m\}_{m=1}^M$ with random admissible values.
Iteration:
1: **while** not convergence **do**
2: Compute $\mathbf{v}^*$ and f^* with $\{d_m\}_{m=1}^M$.
3: Compute $\frac{\partial J}{\partial d_m}$ and $\mathbf{D}$.
4: $\mathbf{d} \leftarrow \mathbf{d} + \gamma\mathbf{D}$, where γ is the step size.
5: **end while**
The final prediction on D_t^u:
$\mathcal{K}_{P_t}f^*(\mathbf{x}_i^t) = \frac{1}{n_l+n_u}\sum_{j=1}^{n_l+n_u} k^t(\mathbf{x}_i^t, \mathbf{x}_j^t)f^*(\mathbf{x}_j^t)$.

Using Eq. (11), $\mathcal{K}_{P_s}f^*(\mathbf{x}_i^s)$ and $\mathcal{K}_{P_t}f^*(\mathbf{x}_i^t)$ can be vectorized as: $\mathcal{K}_{P_s}f^*(\mathbf{x}_i^s) = \sum_{m=1}^M d_m\mathbf{k}_i^s\mathbf{K}_m\mathbf{v}^*$ and $\mathcal{K}_{P_t}f^*(\mathbf{x}_i^t) = \sum_{m=1}^M d_m\mathbf{k}_i^t\mathbf{K}_m\mathbf{v}^*$, where $\mathbf{k}_i^s = 1/n_s[(\mathbf{K}^s)_{i,1}, ..., (\mathbf{K}^s)_{i,n_s}]^T$ and $\mathbf{k}_i^t = 1/(n_l+n_u)[(\mathbf{K})_{i,1}^t, ..., (\mathbf{K}^t)_{i,n_l+n_u}]^T$. As the optimal objective value of Eq. (10), function $J(\mathbf{d})$ is equal to the following expression:

$$\begin{aligned}&\beta\sum_{m=1}^{M} d_m(\mathbf{v}^*)^T\mathbf{K}_m\mathbf{v}^* + \frac{\lambda_s}{n_s}\sum_{i=1}^{n_s}(\sum_{m=1}^{M} d_m\mathbf{k}_i^s\mathbf{K}_m\mathbf{v}^* - y_i^s)^2 \\ &+ \frac{\lambda_t}{n_l}\sum_{i=1}^{n_l}(\sum_{m=1}^{M} d_m\mathbf{k}_i^t\mathbf{K}_m\mathbf{v}^* - y_i^t)^2 + \theta\mathbf{d}^T\mathbf{p}\mathbf{p}^T\mathbf{d}.\end{aligned} \qquad (13)$$

Proposition 2. *$J(\cdot)$ is differentiable and $\frac{\partial J}{\partial d_m}$ can be calculated by the differentiation of Eq. (13) with respect to d_m.*

Proof. The closed-form solution of f^* ensures its unicity for any admissible value of $\mathbf{d}$. Following Theorem 4.1 in (Bonnaus and Shaoiro 1998), $J(\cdot)$ can be proven to be differentiable based on the unicity of f^*. Furthermore, the theorem enables to calculate the derivative $\frac{\partial J}{\partial d_m}$ by the direct differentiation of Eq. (13) with respect to d_m. □

The Reduced Gradient Algorithm The existence and the computation of the gradient of $J(\cdot)$ have been discussed. For solving Problem (9), we propose an efficient and effective procedure which performs the reduced gradient descent on the differentiable $J(\cdot)$. This procedure does converge to the local minimum of $J(\cdot)$ (Luenberger 1984). Specifically, when the gradient is obtained, $\mathbf{d}$ is updated in the descent direction $\mathbf{D}$ computed using Eq. (12) in (Rakotomamonjy et al. 2008). Meanwhile, $\mathbf{D}$ ensures that the constraints $\{\mathbf{d}|\sum_m d_m = 1, d_m > 0\}$ are satisfied. The procedure is summarized in Algorithm 1 with $O(T_{max} \times (n_s + n_l + n_u)^3)$ complexity, where T_{max} is the iteration number. It has been shown that this update scheme leads to rapid convergence (Rakotomamonjy et al. 2008; Wang et al. 2015).

Experiments

Experiment Setup

TFMKL is systematically compared with: 1) SVM trained on the union of labeled source and labeled target data (SVM-st); 2) SVM trained on labeled target data (SVM-t); 3) the

kernel prediction framework which simply combines (labeled) source data and (labeled and unlabeled) target data in the Fredholm integral operator by ignoring distribution difference (Fred-st) (Que, Belkin, and Wan 2014); 4) the kernel prediction framework trained on (labeled and unlabeled) target data in the Fredholm integral operator (Fred-t); 5) unsupervised Kernel Mean Matching (KMM) (Huang et al. 2006); 6) unsupervised Geodesic Flow Kernel (GFK) (Gong et al. 2012) + 1NN; 7) supervised Maximum Margin Domain Transform (MMDT) (Hoffman et al. 2014) and 8) semi-supervised Domain Transfer Multiple Kernel Learning (DTMKL-f) (Duan, Tsang, and Xu 2012).

We evaluate all the methods by empirically searching the parameter space, and the best results are reported. For SVM-st, SVM-t, DTMKL-f and KMM, we choose $C \in \{0.001, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100\}$. For Fred-st, Fred-t and TFMKL, β is searched in the range $[10^{-7}, 10^{1}]$. For DTMKL-f, we search θ, λ and ζ in the range $[10^{-2}, 10^{2}]$. The parameters λ_s, λ_t and θ in TFMKL are searched in the range $[10^{-1}, 10^{1}]$. Across the experiments, TFMKL is found to be robust to these parameters. Base kernels are predetermined as: linear kernel, polynomial kernels with six degrees (i.e., 1.5,1.6,...,2.0) and Gaussian kernels with six bandwidths (i.e., 0.0001, 0.001, 0.01, 0.1, 1, 10). The comparative methods (except TFMKL and DTMKL-f) are evaluated with these 13 base kernels respectively and the best results are reported.

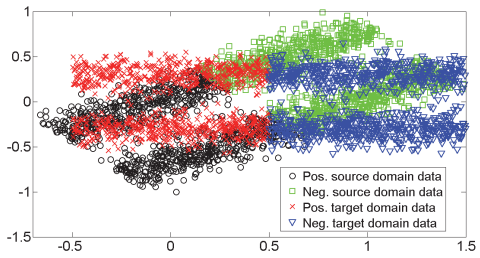

Figure 1: Synthetic example.

Synthetic Example

We construct a synthetic example to illustrate the noise-suppression power of TFMKL in semi-supervised domain adaptation, shown in Figure 1. The two-dimensional synthetic data violates the cluster assumption where multivariate Gaussian noise with variance $\sigma = 0.01$ is added. TFMKL is compared with the semi-supervised DTMKL-f, as well as the standard SVM-st and Fred-st. For each class, 3 labeled target samples are selected. We conduct 50 random selections and the average classification results on the other unlabeled target data are reported in Table 1. As can be seen, DTMKL-f and TFMKL outperform SVM-st and Fred-st due to the exploration of distribution inconsistency. Nevertheless, although DTMKL-f also uses unlabeled target points during classifier learning, it gains limited discriminative information from these noisy data. By contrast, TFMKL provides superior robustness and accuracy, demonstrating its effectiveness in noise resiliency and knowledge transfer.

Table 1: Classification accuracy and standard error.

SVM-st	Fred-st	DTMKL-f	TFMKL
87.86±0.02	90.95±0.01	91.07±0.01	96.02 ±0.11

Cross-Domain Object Recognition

Data Preparation Amazon, DSLR, Webcam and Caltech-256 are four benchmark databases widely used for visual domain adaptation evaluation (Gong et al. 2012). The 10 classes common to these four image databases are extracted, yielding 2,533 images in total. Each database is regarded as a domain and 12 cross-domain image data sets are constructed: $A \rightarrow D$, $A \rightarrow W$, $A \rightarrow C$, $D \rightarrow A$, $D \rightarrow W$, $D \rightarrow C$,...,$C \rightarrow W$. SURF features are extracted and quantized to 800-bin histograms with the codebook trained from a subset of Amazon images.

Results of Visual Object Recognition Following (Hoffman et al. 2014): the source data contains 20 examples per class randomly selected from Amazon source (8 from other source domains); the labeled target data contains 3 labeled examples per class from target domain. Gaussian kernel with the bandwidth 0.001 is employed as outside kernel in Fredholm integral. The averaged classification results on the other unlabeled target data over 20 random splits are shown in Table 2 and some observations can be drawn.

First, SVM-st (or Fred-st) shows higher accuracies than SVM-t (or Fred-t) on $A \rightarrow C$, $D \rightarrow W$, $D \rightarrow C$, $W \rightarrow D$, $W \rightarrow C$ and $C \rightarrow A$. The possible explanation is that under the following two cases, direct involving the source data may be helpful for recognition in the target domain: a) the two domains have some similarities and the distributions may overlap between each other, e.g., DSLR and Webcam, Amazon and Caltech-256; b) the target domain contains relatively rich images and has an extensive distribution in the feature space, e.g., Caltech-256. Second, KMM, GFK, MMDT and DTMKL-f generally outperform SVM-st and SVM-t, verifying the advantage of bridging the gap between domains in domain adaptation. Nevertheless, their results are slightly worse than those of Fred-st and Fred-t. The reason is that noise suppression is exactly required in these challenging visual recognition tasks which contain a great number of noisy images. However, SVM, 1NN and SVR, which are the customized classifiers of these transfer methods, show their limitation in noise suppression. Third, TFMKL performs impressively better than all other methods on most of the data sets (8 out of 12). The average accuracy of TFMKL on the 12 data sets is 57.0%, equivalent to a 4.1% improvement compared to the most competitive Fred-st. Note that Fred-st yields higher accuracies than TFMKL on $D \rightarrow W$ and $W \rightarrow D$, since the two domains are significantly similar. These results show the effectiveness and the robustness of TFMKL, indicating it can successfully minimize distribution mismatch and make confident predictions.

Cross-Domain Text Classification

Data Preparation 20-Newsgroups is a benchmark text corpora organized in a hierarchical structure with differ-

Table 2: Classification accuracy and standard error on the 12 cross-domain object categorization data sets. The best recognition rates are in red and bold font. The second best recognition rates are in blue and italics font.

Data Set	Standard Learning				Transfer Learning				
	SVM-st	SVM-t	Fred-st	Fred-t	DTMKL-f	MMDT	KMM	GFK	TFMKL
$A \rightarrow D$	45.6±0.7	55.9±0.8	46.4±0.8	*58.2±0.9*	46.4±0.7	56.7±1.3	49.0±0.7	50.7±0.8	**62.1±0.9**
$A \rightarrow W$	46.3±0.7	62.4±0.9	50.3±0.9	*69.2±1.1*	48.6±0.8	64.6±1.2	47.4±0.9	58.6±1.0	**69.8±0.9**
$A \rightarrow C$	39.8±0.3	32.0±0.8	39.0±0.4	35.3±0.9	40.5±0.4	36.4±0.8	*40.8±0.3*	36.0±0.5	**43.6±0.6**
$D \rightarrow A$	41.6±0.4	45.7±0.9	51.5±0.5	*51.7±0.8*	42.6±0.4	46.9±1.0	42.6±0.4	45.7±0.6	**52.9±0.7**
$D \rightarrow W$	77.4±0.6	62.1±0.8	**82.4±0.3**	65.4±0.8	76.1±0.5	74.1±0.8	*78.5±0.6*	76.5±0.5	77.8±0.6
$D \rightarrow C$	35.9±0.4	31.7±0.6	*37.7±0.4*	35.2±0.8	37.5±0.5	34.1±0.8	36.9±0.4	32.9±0.5	**37.9±0.4**
$W \rightarrow A$	43.4±0.3	45.6±0.7	49.5±0.5	*52.3±0.7*	45.3±0.4	47.7±0.9	44.4±0.3	44.1±0.4	**54.4±0.4**
$W \rightarrow D$	69.5±0.8	55.1±0.8	**73.2±0.6**	58.2±0.8	69.9±1.1	67.0±1.1	70.4±0.8	*70.5±0.7*	67.7±0.9
$W \rightarrow C$	36.4±0.4	30.4±0.7	37.5±0.3	34.6±0.9	**37.8±0.4**	32.2±0.8	*37.6±0.4*	31.1±0.6	36.1±0.8
$C \rightarrow A$	46.9±0.6	45.3±0.9	*52.1±0.6*	49.9±1.0	49.5±0.9	49.4±0.8	48.0±0.6	44.7±0.8	**54.2±0.9**
$C \rightarrow D$	52.0±1.0	55.8±0.9	55.7±0.9	57.2±1.1	53.1±0.9	56.5±0.9	53.0±1.0	*57.7±1.1*	**59.2±1.1**
$C \rightarrow W$	53.6±0.9	60.3±1.0	60.6±1.2	62.0±1.1	55.4±1.1	*63.8±1.1*	54.6±0.9	63.7±0.8	**68.2±0.9**
Mean	49.0±0.6	48.5±0.8	*52.9±0.6*	52.3±0.8	50.2±0.7	52.5±1.0	50.3±0.6	51.0±0.7	**57.0±0.8**

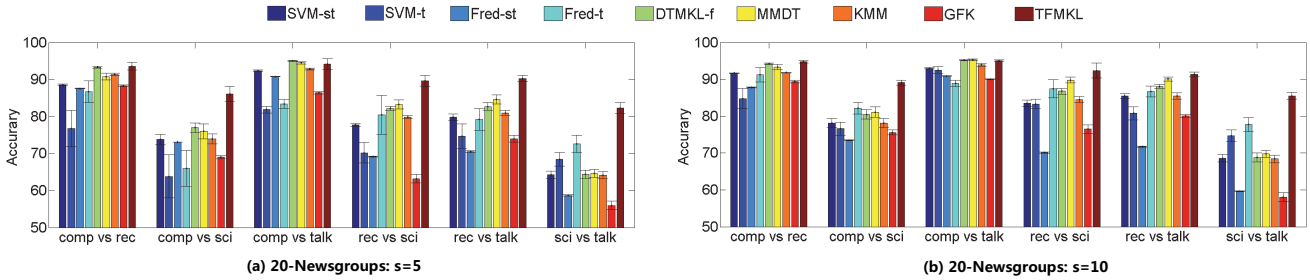

Figure 2: Classification accuracy and standard error: (a) 20-Newsgroups with $s = 5$; (b) 20-Newsgroups with $s = 10$.

ent categories and subcategories (Pan et al. 2011). Data from different subcategories under the same category is related, making the corpora well-suited for constructing cross-domain data sets. As in (Duan, Tsang, and Xu 2012), the four largest main categories in 20-Newsgroups (i.e., comp, rec, sci and talk) are selected to construct the following six cross-domain data sets: *comp vs rec*, *comp vs sci*, *comp vs talk*, *rec vs sci*, *rec vs talk* and *sci vs talk*.

Results of Text Classification Following the setup in (Duan, Tsang, and Xu 2012): the source data contains all the labeled samples in source domain; the labeled target data contains $2s$ labeled examples (s positive and s negative samples) randomly selected from target domain and the unlabeled target data contains the remaining unlabeled examples for training and testing. Linear kernel is used as outside kernel. The classification results are shown in Figure 2 by averaging over 5 random splits with $s = 5$ or $s = 10$ and we have the following observations. (1) The supervised and the semi-supervised transfer methods (MMDT and DTMKL-f) are better than the unsupervised methods (KMM and GFK) in terms of accuracy, verifying the effectiveness of utilizing labeled target data. (2) MMDT and DTMKL-f generally outperform the standard SVM-st, SVM-t, Fred-st and Fred-t on all the data sets except *sci vs talk*, which validates the necessity of domain adaptation. As can be seen, SVM-t (or Fred-t) achieves higher accuracy than SVM-st (or Fred-st) on *sci vs talk*. It means that the two domains have significantly varied distributions, which challenges the existing transfer methods. (3) TFMKL performs consistently better than all other methods, especially when $s = 5$. These results illustrate the superiority of TFMKL in handing multiple difficult cases (e.g., significantly varied distributions and a small amount of labeled target data) for domain adaptation.

Conclusion

In this paper, we have proposed a novel Transfer Fredholm Multiple Kernel Learning (TFMKL) approach to introduce a new paradigm in semi-supervised domain adaptation. Specifically, TFMKL harnesses unlabeled target samples in both the alignment of distributions and the adaptation of predictive models. Such a paradigm is of great importance in tackling challenging cross-domain problems, since it possesses the following advantages: 1) noise resiliency resulting from a good approximation to the "true" data space; 2) capacity of propagating complex knowledge by minimizing distribution difference; 3) flexibility due to the exploration of a unified RKHS from multiple base kernels. For the convergence guarantee, we propose a simple but effective optimization procedure based on the differentiability proof. Comprehensive experimental results validate the advantages of the proposed method.

Acknowledgements

This work is supported by Natural Science Foundation of China (61303164,61402447,61502466,61672501,61602453

and 61503365). This work is also sponsored by Development Plan of Outstanding Young Talent from Institute of Software, Chinese Academy of Science (ISCAS2014-JQ02).

References

Aytar, Y., and Zisserman, A. 2011. Tabula Rasa: Model Transfer for Object Category Detection. In *Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2252-2259.

Bach, F. R.; Lanckriet, G. R. G; and Jordan, M. I. 2004. Multiple Kernel Learning, Conic Duality, and the SMO Algorithm. In *Proceedings of the 21st International Conference on Machine learning* (ICML), 6-13.

Belkin, M.; Niyogi, P.; and Sindhwani, V. 2006. Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. *Journal of Machine Learning Research* 7:2399-2434.

Bonnaus, J. F., and Shaoiro, A. 1998. Optimization Problems with Perturbation: A Guided Tour. *SIAM Review* 40(2):202-227.

Borgwardt, K. M.; Gretton, A.; Rasch, M. J.; Kriegel, H.-P.; Scholkopf, B.; and Smola, A. J. 2006. Integrating Structured Biological Data by Kernel Maximum Mean Discrepancy. *Bioinformatics* 22(14): 49-57.

Boyd, S., and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

Chapelle, O.; Weston, J.; and Schölkopf, B. 2003. Cluster Kernels for Semi-Supervised Learning. In *Proceedings of the 17th Annual Conference on Advances in Neural Information Processing Systems* (NIPS), 585-592.

Chen, M.; Weinberger, K. Q.; and Blitzer, J. C. 2011. Co-Training for Domain Adaptation. In *Proceedings of the 25th Annual Conference on Advances in Neural Information Processing Systems* (NIPS), 2456-2464.

Daumé III, H. 2007. Frustratingly Easy Domain Adaptation. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics* (ACL), 256-263.

Daumé III, H., and Marcu, D. 2006. Domain Adaptation for Statistical Classifiers. *Journal of Artificial Intelligence Research* 26:101-126, 2006.

Duan, L.; Tsang, I. W.; and Xu, D. 2012. Domain Transfer Multiple Kernel Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(3):465-479.

Evgeniou, T.; Micchelli, C. A.; and Pontil, M. 2005. Learning Multiple Tasks with Kernel Methods. *Journal of Machine Learning Research* 6: 615-637.

Gong, B.; Shi, Y.; Sha, F.; and Grauman, K. 2012. Geodesic Flow Kernel for Unsupervised Domain Adaptation. In *Proceedings of the 25th IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2066-2073.

Hoffman, J.; Rodner, E.; Donahue, J.; Kulis, B.; and Saenko, K. 2014. Asymmetric and Category Invariant Feature Transformations for Domain Adaptation. *International Journal of Computer Vision* 109(1):28-41.

Huang, J.; Smola, A. J.; Gretton, A.; Borgwardt, K. M.; and Schölkopf, B. 2006. Correcting Sample Selection Bias by Unlabeled Data. In *Proceedings of the 20th Annual Conference on Advances in Neural Information Processing Systems*, (NIPS) 601-608.

Li, L.; Jin, X.; and Long, M. 2012. Topic Correlation Analysis for Cross-Domain Text Classification. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, (AAAI) 998-1004.

Liao, X.; Xue, Y.; and Carin, L. 2005. Logistic Regression with An Auxiliary Data Source. In *Proceedings of the 22nd International Conference on Machine learning*, (ICML) 505-512.

Luenberger, D. G. 1984. *Linear and Nonlinear Programming*. Addison-Wesley.

Pan, S. J., and Yang, Q. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* 22:1345-1359.

Pan, S. J.; Tsang, I. W.; Kwok, J. T.; and Yang, Q. 2011. Domain Adaptation via Transfer Component Analysis. *IEEE Transactions on Neural Networks* 22(2):199-210.

Que, Q; Belkin, M; and Wan, Y. 2014. Learning with Fredholm Kernels. In *Proceedings of the 28th Annual Conference on Advances in Neural Information Processing Systems* (NIPS), 2951-2959.

Rakotomamonjy, A.; Bach, F. R.; Canu, S.; and Grandvalet, Y. 2008. SimpleMKL. *Journal of Machine Learning Research* 9:2491-2521.

Rifkin, R.; Yeo, G.; and Poggio, T. 2003. Regularized Least Squares Classification. *Advances in Learning Theory: Methods, Model and Applications. NATO Science Series III: Computer and Systems Sciences* 190:131-153.

Scholkopf, B., and Smola, A. J. 2001. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press Cambridge.

Shawe-Taylor, J., and Cristianini, N. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press.

Sun, Q.; Chattopadhyay, R.; Panchanathan, S.; and Ye, J. 2011. A Two-Stage Weighting Framework for Multi-Source Domain Adaptation. In *Proceedings of the 25th Annual Conference on Advances in Neural Information Processing Systems* (NIPS), 505-513.

Wang, X.; Huang, T.; and Schneider, J. 2014. Active Transfer Learning under Model Shift. In *Proceedings of the 31st International Conference on Machine learning* (ICML), 1305-1313.

Wang, W.; Wang, H.; Zhang, C.; and Xu, F. J. 2015. Transfer Feature Representation via Multiple Kernel Learning. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence* (AAAI), 3073-3079.

Yen, E. H.; Lin, T. W; Lin, S. D.; Ravikumar, P. K.; and Dhillon, I. S. 2014. Sparse Random Feature Algorithm as Coordinate Descent in Hilbert Space. In *Proceedings of the 28th Annual Conference on Advances in Neural Information Processing Systems* (NIPS), 2456-2464.