Semi-supervised Subspace Co-Projection for Multi-class Heterogeneous Domain Adaptation

Min Xiao and Yuhong Guo(✉)

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

Abstract. Heterogeneous domain adaptation aims to exploit labeled training data from a source domain for learning prediction models in a target domain under the condition that the two domains have different input feature representation spaces. In this paper, we propose a novel semi-supervised subspace co-projection method to address multi-class heterogeneous domain adaptation. The proposed method projects the instances of the two domains into a co-located latent subspace to bridge the feature divergence gap across domains, while simultaneously training prediction models in the co-projected representation space with labeled training instances from both domains. It also exploits the unlabeled data to promote the consistency of co-projected subspaces from the two domains based on a maximum mean discrepancy criterion. Moreover, to increase the stability and discriminative informativeness of the subspace co-projection, we further exploit the error-correcting output code schemes to incorporate more binary prediction tasks shared across domains into the learning process. We formulate this semi-supervised learning process as a non-convex joint minimization problem and develop an alternating optimization algorithm to solve it. To investigate the empirical performance of the proposed approach, we conduct experiments on cross-lingual text classification and cross-domain digit image classification tasks with heterogeneous feature spaces. The experimental results demonstrate the efficacy of the proposed method on these heterogeneous domain adaptation problems.

1 Introduction

Domain adaptation is the task of exploiting labeled training data in a *label-rich source* domain to train prediction models in a *label-scarce target* domain, aiming to greatly reduce the manual annotation effort in the target domain. Recently, heterogeneous domain adaptation, which generalizes the standard domain adaptation into a more challenging scenario where the source domain and the target domain have different feature spaces, has attracted a lot attention in the research community [6,10,16]. Heterogeneous domain adaptation techniques have applications in many different areas, including image classification in computer vision

A. Appice et al. (Eds.): ECML PKDD 2015, Part II, LNAI 9285, pp. 525–540, 2015.
DOI: 10.1007/978-3-319-23525-7_32

[10,16], drug efficiency prediction in biotechnology [16], cross-language text classification [6] and cross-lingual text retrieval [17] in natural language processing.

A fundamental challenge in heterogeneous domain adaptation lies in the disjoint feature representation spaces of the two domains; with the disjoint feature spaces, a prediction model trained in the source domain cannot be applied in the target domain. A number of representation learning methods have been developed in the literature to address this challenge, including the instance projection methods [6,16] which project instances in the two domains into a common feature space, and the instance transformation methods [10,12] which transform instances from one domain into the other one. These methods however conduct representation learning either in a fully unsupervised manner [16] without exploiting the label information, or in a fully supervised manner [6,10,12] without exploiting the available unlabeled instances. Moreover, some works [16,18] perform representation learning and prediction model training separately, leading to non-optimal representations for the target classification task.

In this paper, we propose a novel semi-supervised subspace co-projection method to address heterogeneous domain adaptation problems, which overcomes the drawbacks of the previous methods mentioned above. The proposed method projects instances in the source and target domains from domain-specific feature spaces to a co-located low-dimensional representation space, while *simultaneously* training prediction models in the projected feature space with labeled instances from the two domains. Moreover, the unlabeled instances are exploited to promote cross-domain instance co-projection by enforcing the empirical mean distributions of the projected source instances and the projected target instances to be similar. Furthermore, we exploit Error-Correcting Output Code (ECOC) schemes [5] to cast a cross-domain multi-class classification task into a large number of cross-domain binary prediction tasks, aiming to increase the stability and discriminative informativeness of the subspace co-projection and enhance cross-domain multi-class classification. The overall semi-supervised learning process is formulated as a joint minimization problem, and solved using an alternating optimization procedure. To evaluate the proposed learning method, we conduct cross-lingual text classification experiments on multilingual Amazon product reviews and cross-domain digit image classification experiments on the UCI handwritten digits data. The experimental results demonstrate the efficacy of the proposed approach for multi-class heterogeneous domain adaptation.

2 Related Work

In this section, we provide a brief review over the related works on heterogeneous domain adaptation, including latent subspace learning methods, instance transformation methods, and auxiliary resources assisted learning methods.

A group of works address heterogeneous domain adaptation by developing latent subspace learning methods that project instances from the domain-specific feature spaces into a common latent subspace [6,13,16,17,20]. In particular, Shi et al. [16] proposed a heterogeneous spectral mapping (HeMap) method,

which learns two projection matrices and projects instances via spectral transformation. Wang et al. [17] proposed a manifold alignment (DAMA) method, which learns projection matrices by using manifold alignment and similarity/dissimilarity constraints constructed on pairs of instances with same/different labels. Duan et al. [6] proposed a heterogeneous feature augmentation (HFA) method, which first projects instances into a common subspace and uses the projected latent features to augment the original features of the instances, and then trains a classification model with the feature-augmented instances. Later, Li et al. [13] extended the HFA method into a semi-supervised HFA (SHFA) method by incorporating unlabeled target training data. Wu et al. [20] proposed to address heterogeneous domain adaptation by performing heterogeneous transfer discriminant analysis of canonical correlations, which maximizes/minimizes the intra/inter-class canonical correlations of the projected instances while simultaneously reducing the data distribution mismatch between the original data and the projected data. Our proposed approach shares similarities with these subspace learning methods on projecting original instances into common representation subspaces. But different from these previous works, our approach exploits both labeled and unlabeled instances and simultaneously learns the projection matrices and the prediction models. Moreover, our approach can naturally exploit error-correcting output code schemes to promote label informative subspace co-projection.

Another group of works developed instance transformation methods to address heterogeneous domain adaptation, which learn asymmetric mapping matrices to transform instances from the source domain to the target domain or vice versa [10,12,18,21]. Kulis et al. [12] proposed an asymmetric regularized cross-domain transformation method that learns an asymmetric feature transformation matrix by performing nonlinear metric learning with similarity/dissimilarity constraints constructed on all pairs of labeled instances. Wang et al. [18] proposed a two-step feature mapping method based on Hilbert-Schmidt Independence Criterion (HSIC) [8] for heterogeneous domain adaptation. It first selects features in each domain based on the HSIC between the instance feature kernel matrix and the instance label kernel matrix, and then maps the selected features across domains based on HSIC. Hoffman et al. [10] proposed a Max-Margin Domain Transforms (MMDT) method to learn domain-invariant image representations. It transforms target instances into the source domain and trains a prediction model in the source domain with the original labeled instances and the transformed labeled instances. Xiao and Guo [21] proposed a semi-supervised kernel matching method for heterogeneous domain adaptation. It learns a prediction function on the labeled source data while mapping the target data points to similar source data points by matching the target kernel matrix to a sub-matrix of the source kernel matrix based on a Hilbert Schmidt Independence Criterion.

In addition to the two groups of methods mentioned above, some other works exploit different types of auxiliary resources to build connections between the source features and the target features, including the ones that use bilingual dictionaries [4,9,19], and the ones that use additional unlabeled image and doc-

uments [22]. However, these auxiliary resource based learning methods are typically designed for specific applications and may have difficulty to be applied on other application tasks.

3 Semi-supervised Multi-class Heterogeneous Domain Adaptation

In this paper, we focus on multi-class heterogeneous domain adaptation problems. We assume in the source domain we have plenty of labeled instances while in the target domain we only have a small number of labeled instances. The two domains have disjoint input feature spaces, $\mathcal{X}_s = \mathbb{R}^{d_s}$ and $\mathcal{X}_t = \mathbb{R}^{d_t}$, where d_s is the dimensionality of the source domain feature space and d_t is the dimensionality of the target domain feature space, but share the same multi-class output label space $\mathcal{Y} = \{-1, 1\}^L$, where L is the number of classes. In particular, let $X_s = [X_s^\ell; X_s^u] \in \mathbb{R}^{n_s \times d_s}$ denote the data matrix in the source domain, where each instance is represented as a row vector. $X_s^\ell \in R^{\ell_s \times d_s}$ is the labeled source data matrix with a corresponding label matrix $Y_s \in \{-1, 1\}^{\ell_s \times L}$, and $X_s^u \in \mathbb{R}^{u_s \times d_s}$ is the unlabeled source data matrix. Each row of the label matrix contains only one positive 1, which indicates the class membership of the corresponding instance. Similarly, let $X_t = [X_t^\ell; X_t^u] \in \mathbb{R}^{n_t \times d_t}$ denote the data matrix in the target domain, where $X_t^\ell \in R^{\ell_t \times d_t}$ is the labeled target data matrix with a corresponding label matrix $Y_t \in \{-1, 1\}^{\ell_t \times L}$ and $X_t^u \in \mathbb{R}^{u_t \times d_t}$ is the unlabeled target data matrix. The number of labeled target domain instances ℓ_t is small and the number of labeled source domain instances ℓ_s is much larger than ℓ_t.

In this section, we present a semi-supervised subspace co-projection method to address heterogeneous multi-class domain adaptation under the setting described above. We formulate a co-projection based discriminative subspace learning method to simultaneously project the instances from both domains into a co-located subspace and train a multi-class classification model in the projected subspace, while exploiting the available unlabeled data to enforce a maximum mean discrepancy criterion across domains in the projected subspace. We further exploit ECOC schemes to enhance the discriminative informativeness of the projected subspace while directly addressing multi-class classification problems.

3.1 Semi-supervised Learning Framework

With the disjoint feature spaces across domains, traditional machine learning methods and homogeneous domain adaptation methods cannot be directly applied in the heterogeneous domain adaptation setting. However, if we can transform the two disjoint feature spaces $\mathcal{X}_s$ and $\mathcal{X}_t$ into a common subspace $\mathcal{Z} = \mathbb{R}^m$ with two transformation functions $\psi_s : \mathcal{X}_s \longrightarrow \mathcal{Z}$ and $\psi_t : \mathcal{X}_t \longrightarrow \mathcal{Z}$, we can then build a unified prediction model in the common subspace to adapt information across domains. Since the same multi-class prediction task is shared across the source domain and the target domain, i.e., the two domains have the

same output label space, we can identify a useful common subspace representation of the data by enforcing the discriminative informativeness of the subspace representation of the labeled data in both domains for the common multi-class prediction task. Based on this motivation, we propose to project the instances from the source domain and the target domain into a common subspace using two projection matrices U_s and U_t respectively such that $\psi_s(X_s) = X_s U_s$ and $\psi_t(X_t) = X_t U_t$, while simultaneously training shared cross-domain prediction models using the projected data. This process can be formulated as the following minimization problem over the projection matrices and the prediction model parameters

$$\begin{aligned}\min_{U_s,U_t,W} \quad & \frac{1}{\ell_s+\beta\ell_t}\mathcal{L}\left(f(X_s^{\ell}U_s,W),\phi(Y_s)\right)+\frac{\alpha_s}{2}R(U_s)+ \\ & \frac{\beta}{\ell_s+\beta\ell_t}\mathcal{L}\left(f(X_t^{\ell}U_t,W),\phi(Y_t)\right)+\frac{\alpha_t}{2}R(U_t)+\frac{\gamma}{2}R(W) \end{aligned} \tag{1}$$

where $U_s \in \mathbb{R}^{d_s \times m}$ and $U_t \in \mathbb{R}^{d_t \times m}$ are two projection matrices that transform the input data in the source domain and target domain respectively to a common and low dimensional feature space, such that $m < \min(d_s, d_t)$; $f(\cdot,\cdot)$ is a prediction function for both domains in the projected common feature space and $W \in \mathbb{R}^{m \times K}$ is the prediction model parameter matrix; $R(\cdot)$ denotes a regularization function; $\phi(\cdot)$ denotes a label transformation function, which transforms the multi-class label vectors from the original space $\{-1,1\}^L$ to a new space $\{-1,1\}^K$; $\mathcal{L}(\cdot,\cdot)$ is a loss function; and $\{\beta,\alpha_s,\alpha_t,\gamma\}$ are trade-off parameters. We introduce the label transformation function $\phi(\cdot)$ to provide a mechanism for incorporating label encoding schemes later.

Since the same prediction model is shared across the two domains, we expect that the discriminative subspace learning framework above can successfully identify a common subspace representation if there are sufficient labeled instances in both domains to enforce the predictive consistency of the subspace projections. However, there are typically only a small number of labeled instances in the target domain, which might lead to poor subspace identification in the target domain. To overcome this potential problem, we further incorporate unlabeled instances to assist the subspace co-projection across domains. Specifically, we assume the empirical marginal instance distributions of the two domains in the projected subspace should be similar, i.e., $P(\psi(X_s))$ and $P(\psi(X_t))$ are similar, and hence the prediction model built in the projected subspace using the labeled source domain instances can work well for the target domain. We thus propose to minimize the distance between the means of the projected instances (both labeled and unlabeled) in the two domains, $\mathcal{D}(\overline{\psi}(X_s),\overline{\psi}(X_t))$. The empirical mean vector $\overline{\psi}(X_s)$ in the source domain can be expressed as $\overline{\psi}(X_s) = \frac{1}{n_s}\mathbf{1}_{n_s}^{\top} X_s U_s$, where $\mathbf{1}_{n_s}$ denotes a column vector of 1s with length n_s. Similarly, the empirical mean vector $\overline{\psi}(X_t)$ in the target domain can be expressed as $\overline{\psi}(X_t) = \frac{1}{n_t}\mathbf{1}_{n_t}^{\top} X_t U_t$, where $\mathbf{1}_{n_t}$ denotes a column vector of 1s with length n_t. By incorporating the empirical mean vector distance measure into our formulation above, we produce the following semi-supervised heterogeneous domain

adaptation framework

$$\min_{U_s,U_t,W} \frac{1}{\ell_s+\beta\ell_t}\mathcal{L}\left(f(X_s^{\ell}U_s,W),\phi(Y_s)\right)+\frac{\alpha_s}{2}R(U_s)+ \frac{\beta}{\ell_s+\beta\ell_t}\mathcal{L}\left(f(X_t^{\ell}U_t,W),\phi(Y_t)\right)+\frac{\alpha_t}{2}R(U_t)+ \frac{\gamma}{2}R(W)+\eta\,\mathcal{D}\Big(\frac{1}{n_s}\mathbf{1}_{n_s}^{\top}X_sU_s,\frac{1}{n_t}\mathbf{1}_{n_t}^{\top}X_tU_t\Big) \tag{2}$$

This framework will ensure the common subspace identified across domains to be informative for the shared prediction model in the two domains, while enforcing the two domains have similar marginal instance distributions in the projected subspace to facilitate information adaptation across domains.

We expect the semi-supervised formulation above to provide a general framework for identifying discriminative common subspace representations for effective information adaptation across domains. Nevertheless, to produce a specific learning problem, we need to consider specific prediction functions, loss functions, regularization functions and distance functions. In this work, we use a linear prediction function $f(x,w)=xw$, a least squares loss function $\mathcal{L}(\hat{y},y)=(\hat{y}-y)^2$, and a squared L2-norm regularization function $R(w)=\|w\|_2^2$. We consider an Euclidean distance function $\mathcal{D}(\cdot,\cdot)$, which leads to a *maximum mean discrepancy criterion* [2]. The maximum mean discrepancy criterion has been used in the literature to induce similar marginal instance distributions across domains in homogeneous domain adaptation setting, and it has been shown to be effective in bridging the domain divergence gaps [3,14]. We expect such an empirical distribution based criterion can be useful for learning the common subspace across heterogeneous domains in our setting. These specific components together lead to the following semi-supervised learning problem

$$\min_{U_s,U_t,W} \frac{1}{\ell_s+\beta\ell_t}\left\|X_s^{\ell}U_sW-\phi(Y_s)\right\|_F^2+\frac{\alpha_s}{2}\|U_s\|_F^2+ \frac{\beta}{\ell_s+\beta\ell_t}\left\|X_t^{\ell}U_tW-\phi(Y_t)\right\|_F^2+\frac{\alpha_t}{2}\|U_t\|_F^2+ \frac{\gamma}{2}\|W\|_F^2+\eta\left\|\frac{1}{n_s}\mathbf{1}_{n_s}^{\top}X_sU_s-\frac{1}{n_t}\mathbf{1}_{n_t}^{\top}X_tU_t\right\|_2^2 \tag{3}$$

where $\|.\|_F$ denotes the Frobenius norm, $\|.\|_2$ denotes the L2 norm, and $\{\alpha_s,\alpha_t,\beta,\gamma,\eta\}$ are trade-off parameters.

The label transformation function $\phi(\cdot)$ allows one to use different multi-class classification schemes within the proposed framework above. For example, if we use the standard one-vs-all (OVA) scheme to address multi-class classification, i.e., training one binary predictor for each label class, we then will have an identical label transformation function $\phi(Y)=Y$, and set $K=L$ for the size of the prediction model parameter matrix W.

3.2 Multi-class Classification with ECOC Schemes

In addition to the one-vs-all (OVA) scheme for multi-class classification, we further exploit the general error-correcting output code (ECOC) [5] schemes for multi-class classification. There are two reasons to use ECOC schemes in our learning framework. First, ECOC schemes have the capacity of encoding a multi-class classification problem into many more binary classification problems than the OVA scheme. More cross-domain binary classification tasks can help to increase the stability and prediction informativeness of the subspace co-projection in the proposed approach above, and lead to more robust domain adaptation performance. Second, ECOC schemes have been used in the literature to robustly solve multi-class classification problems with good empirical results [5]. Incorporating an ECOC scheme in our learning framework will benefit our multi-class classification task.

An ECOC scheme has two components: encoding process and decoding process. Given a L-class classification problem, in the encoding process, an ECOC scheme assigns a codeword from $\{-1, +1\}^K$ to each of the L classes, where K is the length of the codeword. All the codewords for the L classes can then form a codeword matrix $M \in \{-1, +1\}^{L \times K}$, whose each row contains the codeword for one of the L classes. Based on such a codeword matrix, the label transformation function $\phi(\cdot)$ can transform any given label vector from the one-vs-all form into a new label vector with length K, while converting the L-class classification problem to K binary classification problems, each of which corresponds to one column of the codeword matrix M. In the decoding process, one can simply compare the predicted codeword with the codewords in the codeword matrix M to determine the predicted class (one of the L classes). In this work, we use the Euclidean distance based loss decoding [7].

There are different ECOC schemes proposed in the literature. One standard scheme is the exhaustive ECOC [5], which constructs codewords with length $K = 2^{L-1} - 1$. Dense random encoding [1] is another simple ECOC encoding scheme. For a given codeword length K, the random encoding constructs the codeword vectors for the L classes by randomly filling the vectors with 1s and -1s, and then selects the codeword matrix with the largest sum of column separation and row separation from the results of multiple random repeats.

4 Training Algorithm

The semi-supervised learning problem in Eq (3) is a non-convex joint minimization problem over the three parameter matrices, U_s, U_t, and W. But the problem is convex in each individual parameter matrix given the other two fixed, and has closed-form solutions.

First, given fixed U_t and W, the optimization problem over U_s in Eq (3) is simply a least squares minimization problem. By setting the derivative of the objective function regarding U_s to zeros, we obtain the following closed-form solution

$$\mathrm{vec}(U_s) = \left((WW^\top) \otimes A_s + I \otimes B_s\right)^{-1} \mathrm{vec}(Q_s) \tag{4}$$

where $\otimes$ denotes the Kronecker product operator, $\mathrm{vec}(\cdot)$ is the matrix vectorization operator, I is an identity matrix with proper size in the given context, and

$$A_s = \frac{2}{\ell_s + \beta\ell_t} X_s^{\ell\top} X_s^{\ell},$$
$$B_s = \alpha_s I + \frac{2\eta}{n_s^2} X_s^{\top} \mathbf{1}_{n_s} \mathbf{1}_{n_s}^{\top} X_s,$$
$$Q_s = \frac{2}{\ell_s + \beta\ell_t} X_s^{\ell\top} \phi(Y_s) W^{\top} + \frac{2\eta}{n_s n_t} X_s^{\top} \mathbf{1}_{n_s} \mathbf{1}_{n_t}^{\top} X_t U_t,$$

Similarly, given fixed U_s and W, the optimization problem over U_t in Eq (3) has the following closed-form solution

$$\mathrm{vec}(U_t) = \left((WW^{\top}) \otimes A_t + I \otimes B_t\right)^{-1} \mathrm{vec}(Q_t) \tag{5}$$

where

$$A_t = \frac{2\beta}{\ell_s + \beta\ell_t} X_t^{\ell\top} X_t^{\ell},$$
$$B_t = \alpha_t I + \frac{2\eta}{n_t^2} X_t^{\top} \mathbf{1}_{n_t} \mathbf{1}_{n_t}^{\top} X_t,$$
$$Q_t = \frac{2\beta}{\ell_s + \beta\ell_t} X_t^{\ell\top} \phi(Y_t) W^{\top} + \frac{2\eta}{n_s n_t} X_t^{\top} \mathbf{1}_{n_t} \mathbf{1}_{n_s}^{\top} X_s U_s.$$

Finally, the optimization problem over W given fixed U_s and U_t has the following closed-form solution

$$W = \left(\frac{2N_x}{\ell_s + \beta\ell_t} + \gamma I\right)^{-1} \left(\frac{2N_y}{\ell_s + \beta\ell_t}\right) \tag{6}$$

where

$$N_x = U_s^{\top} X_s^{\ell\top} X_s^{\ell} U_s + \beta U_t^{\top} X_t^{\ell\top} X_t^{\ell} U_t,$$
$$N_y = U_s^{\top} X_s^{\ell\top} \phi(Y_s) + \beta U_t^{\top} X_t^{\ell\top} \phi(Y_t).$$

Given these closed-form solutions for each individual subproblem, we use an alternating procedure to solve the optimization problem in Eq (3) in an iterative manner. After a random initialization over $\{U_s, U_t, W\}$, in each iteration the alternating procedure sequentially updates U_s, U_t and W according to equations (4), (5) and (6) respectively to minimize the objective function. We stop the iteration until a local optimal objective has been reached. On high-dimensional data, where the closed-form solutions in (4) and (5) involve large matrix inversions, we use a conjugate gradient descent algorithm to solve the subproblems over U_s and U_t to achieve scalability.

5 Experiments

We conducted experiments on cross-lingual text classification tasks and digit image classification tasks with heterogeneous feature spaces. In this section we report the experimental settings and the empirical results.

5.1 Datasets and Methods

We conducted experiments on two types of data, text data and image data, using Amazon product reviews [15] and UCI handwritten digits [11] respectively. The Amazon product review dataset is a multilingual sentiment classification dataset. It contains reviews from three different categories (Books, DVD and Music), written in four different languages (*English (E)* , *French (F)* , *German (G)* and *Japanese (J)*), where each review is represented as a term-frequency feature vector. With this dataset, we constructed 12 cross-lingual multi-class classification tasks with the three categories $\{Books, DVD, Music\}$ as classes, one for each source-target language pair. For example, the task *E2F* uses *English* as the source language and *French* as the target language. For each task, there are 4000 views for each class in each language domain.

The UCI handwritten digits dataset contains 2000 digit images, evenly distributed among ten digit classes (from zero to nine). We randomly split the dataset into two subsets with equal size as two domains. Images in one domain are represented using the feature set of the Zernike moments (Zer), while images in the other domain are represented using the feature set of the profile correlations (Fac). We then constructed two heterogeneous domain adaptation tasks, *Fac2Zer* and *Zer2Fac*, one for each ordered source-target domain pair.

Methods: For each constructed heterogeneous domain adaptation task, we compared the following methods: (1) *TB* - this is a target baseline method that trains a classifier using only the labeled instances in the target domain. (2) *HeMap* - this is an unsupervised representation learning method for heterogeneous domain adaptation [16], which first learns two projection matrices for the two domains and then trains a classifier using the projected labeled instances from the two domains. (3) *DAMA* - this is a semi-supervised heterogeneous domain adaptation method proposed in [17], which performs representation learning and model training in separate steps. (4) *MMDT* - this is a maximum margin domain transform method for heterogeneous domain adaptation [10]. (5) *SHFA* - this is a semi-supervised heterogeneous feature augmentation-based domain adaptation method [13]. (6) *SCP-OVA* - this is the proposed subspace co-projection method with the one-verse-all (OVA) scheme for multi-class classification. (7) *SCP-ECOC* - this is the proposed subspace co-projection method with the exhaustive ECOC scheme for multi-class classification. The DAMA method [17] cannot handle the original high-dimensional features of the review data, we thus applied PCA to reduce the dimensionality of the input features in each language domain to 1000, as suggested in the *SHFA* work [13]. The alternating training algorithm for our proposed approaches is very efficient, and it typically converges within 30 iterations in our experiments.

5.2 Cross-lingual Text Classification

For each of the 12 cross-lingual multi-class classification tasks on Amazon product reviews, there are 4000 instances for each of the three classes in each domain.

Table 1. Average test accuracy (± standard deviations) (%) over 10 runs for cross-lingual text classification tasks.

TASK	TB	HeMap	DAMA	MMDT	SHFA	SCP-OVA	SCP-ECOC
E2F	73.8±0.5	73.8±0.4	74.2±0.5	78.2±0.5	78.4±0.4	79.2±0.5	**80.6**±0.4
E2G	72.4±0.5	76.5±0.5	77.0±0.4	79.2±0.4	79.4±0.4	81.0±0.4	**82.2**±0.3
E2J	66.8±0.5	67.3±0.5	67.6±0.5	72.7±0.5	70.6±0.8	73.4±0.6	**74.4**±0.6
F2E	72.8±0.6	79.3±0.6	80.3±0.5	82.2±0.4	82.4±0.4	84.3±0.3	**85.6**±0.2
F2G	72.4±0.5	76.3±0.4	77.7±0.6	79.4±0.4	79.5±0.4	80.9±0.4	**82.2**±0.3
F2J	66.8±0.5	67.9±0.8	68.4±0.4	72.6±0.5	70.5±0.8	73.4±0.7	**74.5**±0.6
G2E	72.8±0.6	79.8±0.4	80.6±0.6	82.2±0.4	82.4±0.4	84.5±0.3	**85.5**±0.2
G2F	73.8±0.5	73.9±0.4	75.0±0.5	78.2±0.5	78.4±0.4	79.4±0.5	**80.6**±0.4
G2J	66.8±0.5	65.8±1.0	67.5±0.6	72.6±0.5	70.5±0.8	73.3±0.7	**74.4**±0.6
J2E	72.8±0.6	81.0±0.4	81.2±0.4	82.2±0.4	82.5±0.5	84.2±0.2	**85.5**±0.2
J2F	73.8±0.5	74.8±0.3	75.1±0.7	78.3±0.5	78.3±0.4	79.3±0.5	**80.5**±0.4
J2G	72.4±0.5	76.4±0.4	77.1±0.6	79.2±0.4	79.3±0.4	81.0±0.4	**82.2**±0.4

We conducted experiments in the following way. In the source domain, we randomly selected 2000 instances from each class as labeled data and used the remaining 2000 instances as unlabeled data. In the target domain, we randomly selected 100 instances and 2900 instances from each class as labeled and unlabeled data respectively. We used all these selected instances for training, and used the remaining 3000 instances (1000 for each class) in the target domain as testing data. For the comparison approaches, *HeMap, DAMA, SCP-OVA, SCP-ECOC*, which involve low dimensional subspaces, we set the dimension of the latent subspaces, m, as 100. Then we performed empirical parameter selection using the first task *E2F* with three runs. For the proposed approaches, *SCP-OVA* and *SCP-ECOC*, we chose α_s and α_t from $\{0.01, 0.1, 1, 10, 100\}$, β from $\{1, 2, 5, 10, 100\}$, η from $\{0.01, 0.1, 1, 10, 100\}$, and chose γ from $\{0.01, 0.1, 1, 10, 100\}$. We picked the parameter setting with the best test classification accuracy for each approach, $\{\alpha_s = 0.1, \alpha_t = 0.1, \beta = 1, \eta = 10, \gamma = 0.1\}$ for *SCP-OVA* and $\{\alpha_s = 10, \alpha_t = 0.1, \beta = 1, \eta = 10, \gamma = 0.1\}$ for *SCP-ECOC*. We conducted parameter selection for the other comparison approaches, *HeMap, DAMA, MMDT, SHFA*, in the same way. Using the selected parameters, for each of the 12 tasks we then repeatedly ran all the comparison methods for 10 times with different random selections of the training instances. The comparison results in terms of average test accuracy in the target domain are reported in Table 1.

From Table 1, we can see that the *TB* baseline method performs poorly across all the twelve tasks, which shows that the 100 labeled target training instances from each class are far from enough to obtain a good classification model in the target language domain. By exploiting the labeled training data from the source language domain, the *HeMap* method improves the prediction performance on most tasks. However, its improvements over *TB* are very small on some tasks and it even performs worse than *TB* on the task *G2J*. The *DAMA* method on the other hand consistently outperforms both *TB* and *HeMap*. The explanation

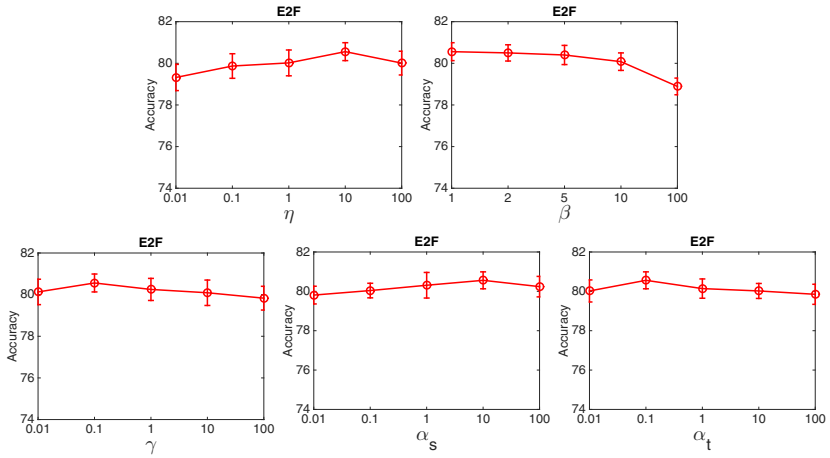

Fig. 1. Parameter sensitivity analysis over trade-off parameters $\{\eta, \beta, \gamma, \alpha_s, \alpha_t\}$.

is that *HeMap* conducts representation learning in a fully unsupervised manner while *DAMA* learns more informative representations in a semi-supervised manner with constraints constructed from the label information. By exploiting the label information *directly* for representation learning and prediction model training, the supervised method *MMDT* and semi-supervised method *SHFA*, further outperform *DAMA* on all the twelve tasks. Nevertheless, our proposed approaches, *SCP-OVA* and *SCP-ECOC*, outperform all the other comparison methods across all the tasks. This suggests that the proposed learning framework, which exploits both labeled and unlabeled training data to simultaneously perform subspace representation learning and prediction model training, is an effective model for heterogeneous domain adaptation. Between the two variants of the proposed model, *SCP-ECOC* consistently outperforms *SCP-OVA* across all the tasks, which suggests that the exhaustive error-correcting output coding is more effective than the one-vs-all coding scheme in our learning framework, while our proposed learning framework has the nice property of naturally incorporating different ECOC schemes.

5.3 Parameter Sensitivity Analysis

Next, we conducted parameter sensitivity analysis for the proposed *SCP-ECOC* approach over the trade-off parameters $\{\eta, \beta, \gamma, \alpha_s, \alpha_t\}$ using the first cross-lingual text classification task, *E2F*. We used the same experimental setting as above, and empirically investigated how the values of the trade-off parameters $\{\eta, \beta, \gamma, \alpha_s, \alpha_t\}$ affect the heterogeneous cross-domain prediction performance. We first conducted sensitivity analysis over η, which controls the relative weight for the mean discrepancy term in the proposed objective function. We conducted experiments with different η values from $\{0.01, 0.1, 1, 10, 100\}$, while fixing the other trade-off parameters as the selected values in the section above. For each η value, we repeated the

Table 2. Average test accuracy (± standard deviations) (%) over 10 runs for digit image classification tasks.

TASK	TB	HeMap	DAMA	MMDT	SHFA	SCP-OVA	SCP-ECOC
Fac2Zer	71.9±0.7	72.0±1.0	72.5±0.6	73.4±1.0	73.8±0.6	75.0±0.8	**76.6**±0.5
Zer2Fac	83.8±0.9	84.2±0.9	85.4±0.6	87.0±1.1	87.6±0.7	88.7±0.7	**90.4**±0.5

experiment 10 times based on random partitions of the dataset and reported the average test performance in the top left figure of Figure 1. We can see *SCP-ECOC* produces the highest test accuracy when η equals 10. As η controls the contribution weight of the maximum mean discrepancy (MMD) criterion across the two domains, the good performance of the large value of η suggests that the MMD term is helpful for improving the cross-domain prediction performance. Another observation is that although the test accuracy varies as we change the value of η, the changes are small and the test accuracies produced by *SCP-ECOC* across the whole range of different η values are all higher than the other comparison methods, *TB, HeMap, DAMA, MMDT* and *SHFA* (see both Figure 1 and Table 1). This suggests that the proposed *SCP-ECOC* is not very sensitive to η within the studied range of values.

We next studied how β affects cross-lingual test classification accuracy. Note that β can be viewed as the relative weight ratio between a labeled target domain instance and a labeled source domain instance regarding their contribution to the training loss. As we have many more labeled training instances in the source domain than in the target domain and we aim to learn a classification model that works well in the target domain, it is reasonable to give a target domain instance larger (or equal) weight than a source domain instance and consider $\beta \geq 1$. In particular, we conducted experiments with different β values from $\{1, 2, 5, 10, 100\}$ while fixing all the other trade-off parameters as the selected values in the previous section. The average test classification results over 10 repeated runs are reported in the top right figure of Figure 1. We can see that the performance of *SCP-ECOC* is quite stable with β values changing from 1 to 10. However, if placing too much weights (*e.g.*, $\beta = 100$) on the target instances, the test performance degrades. These results suggest that the performance of the proposed *SCP-ECOC* is quite robust to β within a range of reasonable values.

We finally investigated the three trade-off parameters $\{\gamma, \alpha_s, \alpha_t\}$ used for the Frobenius norm regularization terms over W, U_s, and U_t respectively. We conducted experiments similarly as above. For each of the three parameters, we repeated the experiment 10 times for each of its values in $\{0.01, 0.1, 1, 10, 100\}$ while fixing all the other trade-off parameters as previously selected values. We reported the average test accuracy results in the bottom three figures of Figure 1 for the three parameters $\{\gamma, \alpha_s, \alpha_t\}$ respectively. We can see although the performance of the proposed *SCP-ECOC* changes with the value change for each of the three parameters, the performance variations are very small. The performance of *SCP-ECOC* is quite robust to the values of $\gamma, \alpha_s, \alpha_t$ within the range of values considered in the experiments.

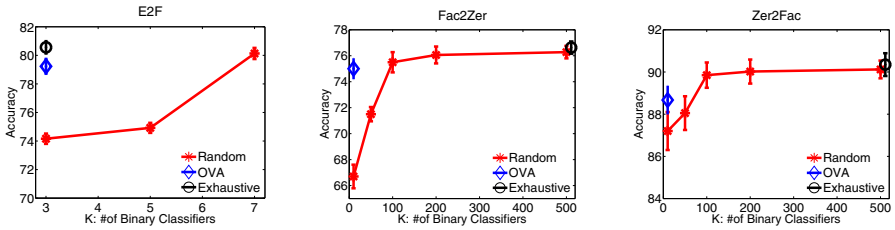

Fig. 2. Empirical comparison of different ECOC schemes.

5.4 Experimental Results on UCI Dataset

We have also conducted experiments using the UCI handwritten digits dataset. The two tasks we constructed on the UCI handwritten digits dataset have different feature spaces across domains, and have 100 instances from each class, i.e., 1000 instances in total, in each domain. For each task, in the source domain, we randomly chose 50 instances from each class (500 in total) as the labeled training data and used the remaining 500 instances as the unlabeled training data. In the target domain, we randomly chose 10 and 70 instances from each class as the labeled and unlabeled training data respectively, and used the remaining instances as the testing instances. For the approaches that involve subspaces, we set the dimension of the subspace as 20. We then used the same parameter selection procedure as before to select values for the trade-off parameters of all the comparison methods using the task *Fac2Zer*. For our proposed approaches, we got $\{\alpha_s = 0.1, \alpha_t = 0.1, \beta = 10, \eta = 0.1, \gamma = 10\}$ for *SCP-OVA* and $\{\alpha_s = 1, \alpha_t = 1, \beta = 1, \eta = 0.1, \gamma = 10\}$ for *SCP-ECOC*. With the selected parameters, for each task, we ran the comparison methods for 10 times with different random selections of the training and testing data. The average test accuracy results are reported in Table 2.

We can see that by exploiting the existing labeled data from the auxiliary source domain, all the heterogeneous domain adaptation methods outperform the baseline method on learning prediction models in the target domain. This again shows the importance of performing heterogeneous domain adaptation. Nevertheless, these few methods used in our experiments also demonstrated different efficacies on heterogeneous domain adaptation. *HeMap* displays similar performance as in the cross-lingual text classification experiments, with limited improvements over the baseline *TB*. The methods *DAMA, SHFA* and *MMDT* outperform *HeMap*, while our proposed two approaches outperform all the other comparison methods. Between the two proposed approaches, again *SCP-ECOC* outperforms *SCP-OVA*. All these results again verified the efficacy of the proposed learning framework.

5.5 Impact of the ECOC Encoding Schemes

We also conducted experiments to further study the influence of different ECOC encoding schemes, especially the different numbers of binary classifiers, on the proposed heterogeneous domain adaptation framework. In particular, we compared the performance of one-vs-all (OVA) scheme, exhaustive ECOC scheme and dense random ECOC encoding schemes [1]. For a L-class classification problem, the OVA scheme transforms the problem into a set of L binary classification problems, the exhaustive ECOC scheme transforms the problem into a set of $(2^{L-1} - 1)$ binary classification problems, while the random ECOC encoding scheme transforms the problem into a given number of K binary classification problems.

We conducted experiments on the first cross-lingual text classification task, *E2F* and the two tasks on UCI digits dataset, *Fac2Zer* and *Zer2Fac*. The *E2F* is a 3-class classification task, and we tested the random encoding ECOC scheme with different K values from $\{3, 5, 7\}$. The *Fac2Zer* and *Zer2Fac* are 10-class classification tasks, and we tested the random encoding ECOC scheme with different K values from $\{10, 50, 100, 200, 500\}$. The experimental results are reported in Figure 2. We can see that the exhaustive ECOC encoding scheme demonstrates the best performance on all the three tasks, even though its codeword length is smaller than the random schemes in some cases on the *E2F* task where the class number is small. This is reasonable since the codeword matrix generated by the exhaustive ECOC scheme typically has much better row and column separations than randomly generated codeword matrix. With the same codeword length, even the OVA scheme produces better performance than the random scheme. But with the increasing of the number of binary classifiers, i.e., the codeword length K, the performance of the proposed approach based on random encoding ECOC improves quickly. In particular, on *Fac2Zer* and *Zer2Fac*, when K increases from 10 to 100, the performance of the proposed approach increases dramatically. Similar performance is observed on *E2F* as well. This observation verifies our hypothesis that incorporating more binary classification tasks can help to increase the stability and usefulness of the subspace co-projection in the proposed learning framework and induce better domain adaptation performance.

6 Conclusion

In this paper, we developed a novel semi-supervised subspace co-projection approach to address multi-class heterogeneous domain adaptation problems, where the source domain and the target domain have disjoint input feature spaces. The proposed method projects instances in the two domains into a co-located latent subspace, while *simultaneously* training prediction models in the projected feature space. It also exploits the unlabeled data to promote the consistency of subspace co-projection from the two domains. Moreover, the proposed learning framework can naturally exploit error-correcting output codes for multi-class classification to enforce the informativeness of the subspace co-projection. We formulated the overall semi-supervised learning process as a joint minimization

problem, and solved it using an alternating optimization procedure. To investigate the empirical performance of the proposed approach, we conducted cross-lingual text classification experiments on the Amazon product reviews and cross-domain image classification experiments on the UCI digits dataset. The empirical results demonstrated the effectiveness of the proposed approach comparing to a number of state-of-the-art heterogeneous domain adaptation methods.

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

References

1. Allwein, E., Schapire, R., Singer, Y.: Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of Machine Learning Research (JMLR) **1**, 113–141 (2001)
2. Borgwardt, K., Gretton, A., Rasch, M., Kriegel, H., Schölkopf, B., Smola, A.: Integrating structured biological data by kernel maximum mean discrepancy. In: Proceedings of the International Conference on Intelligent Systems for Molecular Biology (2006)
3. Chattopadhyay, R., Fan, W., Davidson, I., Panchanathan, S., Ye, J.: Joint transfer and batch-mode active learning. In: Proceedings of the International Conference on Machine Learning (ICML) (2013)
4. Dai, W., Chen, Y., Xue, G., Yang, Q., Yu, Y.: Translated learning: transfer learning across different feature spaces. In: Advances in Neural Information Processing Systems (NIPS) (2008)
5. Dietterich, T., Bakiri, G.: Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Interlligence Research (JAIR) **2**(1), 263–286 (1995)
6. Duan, L., Xu, D., Tsang, I.: Learning with augmented features for heterogeneous domain adaptation. In: Proceedings of the International Conference on Machine Learning (ICML) (2012)
7. Escalera, S., Pujol, O., Radeva, P.: On the decoding process in ternary error-correcting output codes. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **32**(1), 120–134 (2010)
8. Gretton, A., Bousquet, O., Smola, A.J., Schölkopf, B.: Measuring statistical dependence with Hilbert-Schmidt norms. In: Jain, S., Simon, H.U., Tomita, E. (eds.) ALT 2005. LNCS (LNAI), vol. 3734, pp. 63–77. Springer, Heidelberg (2005)
9. He, J., Liu, Y., Yang, Q.: Linking heterogeneous input spaces with pivots for multi-task learning. In: Proceedings of SIAM International Conference on Data Mining (SDM) (2014)
10. Hoffman, J., Rodner, E., Donahue, J., Darrell, T., Saenko, K.: Efficient learning of domain-invariant image representations. In: Proceedings of the International Conference on Learning Representations (ICLR) (2013)
11. Jain, A., Duin, R., Mao, J.: Statistical pattern recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **22**(1), 4–37 (2000)
12. Kulis, B., Saenko, K., Darrell, T.: What you saw is not what you get: domain adaptation using asymmetric kernel transforms. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2011)

13. Li, W., Duan, L., Xu, D., Tsang, I.: Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **36**(6), 1134–1148 (2014)
14. Pan, S., Tsang, I., Kwok, J., Yang, Q.: Domain adaptation via transfer component analysis. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) (2009)
15. Prettenhofer, P., Stein, B.: Cross-language text classification using structural correspondence learning. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (2010)
16. Shi, X., Liu, Q., Fan, W., Yu, P., Zhu, R.: Transfer learning on heterogenous feature spaces via spectral transformation. In: Proceedings of the IEEE International Conference on Data Mining (ICDM) (2010)
17. Wang, C., Mahadevan, S.: Heterogeneous domain adaptation using manifold alignment. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) (2011)
18. Wang, H., Yang, Q.: Transfer learning by structural analogy. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2011)
19. Wei, B., Pal, C.: Heterogeneous transfer learning with rbms. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2011)
20. Wu, X., Wang, H., Liu, C., Jia, Y.: Cross-view action recognition over heterogeneous feature spaces. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2013)
21. Xiao, M., Guo, Y.: Feature space independent semi-supervised domain adaptation via kernel matching. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **37**(1), 54–66 (2014)
22. Zhu, Y., Chen, Y., Lu, Z., Pan, S., Xue, G., Yu, Y., Yang, Q.: Heterogeneous transfer learning for image classification. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2012)