Domain Transfer Dimensionality Reduction via Discriminant Kernel Learning

Ming Zeng and Jiangtao Ren

Sun Yat-Sen University, Guangzhou, 510006, China
mingtsang.zm@gmail.com,
issrjt@mail.sysu.edu.cn

Abstract. Kernel discriminant analysis (KDA) is a popular technique for discriminative dimensionality reduction in data analysis. But, when a limited number of labeled data is available, it is often hard to extract the required low dimensional representation from a high dimensional feature space. Thus, one expects to improve the performance with the labeled data in other domains. In this paper, we propose a method, referred to as the domain transfer discriminant kernel learning (DTDKL), to find the optimal kernel by using the other labeled data from out-of-domain distribution to carry out discriminant dimensionality reduction. Our method learns a kernel function and discriminative projection by maximizing the Fisher discriminant distance and minimizing the mismatch between the in-domain and out-of-domain distributions simultaneously, by which we may get a better feature space for discriminative dimensionality reduction with cross-domain.

Keywords: Discriminant Kernel Learning, Dimensionality Reduction, Transfer Learning.

1 Introduction

In many real-world applications, such as image processing, computational biology and natural language processing, the dimensionality of data is usually very high. Due to the complexity and noise of high-dimensional data, the effectiveness of regression or classification is limited. This can be improved via dimensionality reduction which finds a compact representation of the data for classification.

A more popular technique for dimensionality reduction is discriminant analysis. To handle nonlinear problems, the kernel discriminant analysis (KDA) is proposed in [1], which computes the discriminative projection from the data set that is mapped nonlinearly into the reproducing kernel Hilbert space (RKHS). We observe that the kernel is chosen before learning in the KDA method. However, the kernel-based learning methods are desirable when integrating the tuning of kernel into the learning space.

In addition, the discriminant multiple kernel learning methods require a plenty of labeled samples to discriminate the unlabeled data from each class. In real-world applications, it is usually costly or even impossible to get such a huge

P.-N. Tan et al. (Eds.): PAKDD 2012, Part II, LNAI 7302, pp. 280–291, 2012.

number of labeled samples from the same distribution. When this situation occurs, the performance of discriminant kernel learning methods is poor. Then one expects to carry out discriminant analysis with the help of other related labeled data from other domains. This brings out the cross-domain problem since the existing discriminant kernel learning makes the assumption that the training data and the test data are independent and identical. To resolve this problem, cross transfer learning is proposed, whose aim is to improve learning in the in-domain by porting the labeled sample from out-of-domain to that from in-domain to carry out dimensionality reduction. Several works have been done by combining unsupervised dimensionality with clustering, such as transferred dimensionality analysis(TDA) [2], which intends to select the most discriminative subspace and clustering at the same time. Maximum mean discrepancy embedding(MMDE)[3] tries to find a subspace where training and test samples distribute similarly to solve the sample selection bias problem in an unsupervised way. S.Si et al.[4] proposed using evolutionary cross-domain discriminative Hessian eigenmaps by minimizing the quadratic distance between the distribution of the training set and that of the test set. However, it could not solve non-linear problems.

In this paper, we develop a new dimensionality reduction method, called domain transfer discriminant kernel learning method (DTDKL), which transfers the knowledge from labeled data in out-of-domain to the in-domain by explicitly carrying out kernel discriminant learning and transfer leaning in a coherent way. More specifically, DTDKL tries to find a projection to maximize the Fisher discriminant ratio in the optimal feature space and minimize the maximum mean discrepancy (MMD) of the different distributions simultaneously. In fact, DTDKL provides a method to learn an optimal kernel function and discriminant projection at the same time.

The key contributions of the paper can be highlighted as follows:

- To the best of our knowledge, DTDKL is the first semi-supervised cross-domain discriminant kernel learning method. In contrast to the prior discriminant kernel learning method, DTDKL does not assume that the training and test data are drawn from the same distribution. Moreover, a novel dimensionality reduction method with cross-domain is proposed, whose objects are to maximize the Fisher discriminant ratio while minimizing the maximum mean discrepancy of different distributions.
- By comparing the state-of-the-art dimensionality reduction methods, DTDKL performs better in the dataset of SyskillWebert, Reuters-21578 and 20-Newsgroup ensuring promising performance in real applications.

The rest of this paper is organized as follows: Section 2 presents the related works and preliminaries of DTDKL; Section 3 proposes DTDKL method by embedding maximum mean discrepancy (MMD) into discriminant analysis to tackle the cross-domain problem; Section 4 presents our experimental results to demonstrate its applications. Finally, we conclude the study in Section 5.

2 Brief Review of Prior Work

2.1 Discriminant Multiple Kernel Learning

Dimensionality reduction has always attracted amount of attention. Various methods have been proposed in a recent survey [5] to solve this problem. The canonical dimensionality reduction algorithm is linear discriminant analysis (LDA) [6], which is finding the most discriminative subspace for different classes in the original space. And with the development of kernel-based methods, kernel discriminant analysis has received a lot of interest for nonlinear problems. The KDA algorithm finds the direction in a feature space, defined by a kernel function, onto which the projections of different classes are well separated [1,7]. Note that the kernel function plays a crucial role in kernel methods, and Lanckriet et al. [8] pioneered the work of multiple kernel learning (MKL) in which the optimal kernel is obtained as a linear combination of pre-determined kernel matrices. Based on ideas of MKL, the kernel-based learning method for discriminant analysis was reformulated as semi-definite programming (SDP) in Kim et al. [9]. Ye et al. [10] improved the efficiency of the problem and extended naturally to the multi-class setting by casting the SDP formulation in quadratically constrained quadratic programming (QCQP) and semi-infinite linear programming (SILP).

2.2 Transfer Learning and Maximum Mean Discrepancy Formulation

Semi-supervised learning aims to make use of unlabeled data in the process of supervised learning and it has also been widely used in many areas related to transfer learning. One of the typical branches is to find criteria to estimate the distance between different distributions. A well-known example is Kullback-Leibler (K-L) divergence. Many criteria are parametric for the reason that an intermediate density estimate is usually required. To avoid parametric estimation, some nonparametric methods are proposed to evaluate the distance between the different distributions of data sets. Maximum Mean Discrepancy (MMD) is a effective nonparametric criterion for comparing distributions based on RKHS [11]. Suppose $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$, and $X = (x_1, \ldots, x_{n_1})$, $Y = (y_1, \ldots, y_{n_2})$ be random variable sets drawn from distributions $\mathcal{P}$ and $\mathcal{Q}$, respectively. The maximum mean discrepancy and its empirical estimate is as follows:

$$\mathrm{MMD}[\mathcal{F}, X, Y] := \sup_{f \in \mathcal{F}} \left(\frac{1}{n_1} \sum_{i=1}^{n_1} f(x_i) - \frac{1}{n_2} \sum_{i=1}^{n_2} f(y_i) \right) \tag{1}$$

The function space $\mathcal{F}$ could be replaced by $\mathcal{H}$ which is a universal RKHS. By the fact that in RKHS, $f(x)$ can be expressed as an inner product via $f(x) = \langle \varphi(x), f \rangle_{\mathcal{H}}$, where $\varphi(x) : \mathcal{X} \to \mathcal{H}$, then one may rewrite MMD as follows:

$$\mathrm{MMD} = \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \varphi(x_i) - \frac{1}{n_2} \sum_{i=1}^{n_2} \varphi(y_i) \right\|_{\mathcal{H}} = \|\mu_1 - \mu_2\|_{\mathcal{H}}$$

In terms of the MMD theory [11], the distance between distributions of two samples is equivalent to the distance between the means of the two samples mapped into a RKHS.

3 Semi-supervised Discriminant Analysis in Cross-Domain

Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \{-1, +1\}$ denote the input space and the output space, respectively. Let $\mathcal{D}^{out} = \{(x_i^{out}, y_i^{out}) \in \mathcal{X} \times \mathcal{Y} : 1 \leq i \leq n^{out}\}$ be the set of out-of-domain data samples with $n^{out} = |\mathcal{D}^{out}|$, and $\mathcal{D}^{in} = \mathcal{D}_l^{in} \cup \mathcal{D}_u^{in}$ be the set of in-domain data samples where $\mathcal{D}_l^{in} = \{(x_i^{in}, y_i^{in}) \in \mathcal{X} \times \mathcal{Y} : 1 \leq i \leq n_l^{in}\}$ with $n_l^{in} = |\mathcal{D}_l^{in}|$, and $\mathcal{D}_u^{in} = \{x_i^{in} \in \mathcal{X} : n_{l+1}^{in} \leq i \leq n^{in}\}$ with $n_u^{in} = |\mathcal{D}_u^{in}|$. Typically, $n_l^{in} \ll n_u^{in}$. Let $\mathcal{P}$ and $\mathcal{Q}$ be the marginal distribution of $\mathcal{D}^{in}$ and $\mathcal{D}^{out}$, respectively. We assume that the n^{out} out-of-domain samples and the n^{in} in-domain samples are drawn independently and identically from a fixed but unknown underlying probability distribution $\mathcal{P}$ and $\mathcal{Q}$, respectively. Our task is to predict the labels $y_{n_l+1}^{in}, \ldots, y_n^{in}$, which corresponds to the inputs $x_{n_l+1}^{in}, \ldots, x_n^{in}$ in the in-domain data set.

3.1 Standard Discriminant Kernel Learning Analysis

The standard kernel discriminant analysis learns the kernel and the direction from the labeled samples in $\mathcal{D}_l^{in}$ in order to project the unlabeled samples in $\mathcal{D}_u^{in}$. Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel function. Then, Mercer's Theorem [12] tells us the kernel function implicitly maps the input space $\mathcal{X}$ to a high-dimensional (possibly infinite) Hilbert space $\mathcal{H}$ equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ through a map $\varphi : K \to \mathcal{H}$:

$$K(x, z) = \langle \varphi(x), \varphi(z) \rangle_{\mathcal{H}}, \quad \forall x, z \in \mathcal{X}.$$

This space is called the feature space, and the mapping is called the feature mapping. They depend on the kernel function K and will be denoted as φ_K and $\mathcal{H}_K$.

Let x_+ and x_- denote the collection of data points from positive and negative classes, respectively. Then the total number of data points in the training set $\mathcal{D}_l^{in}$ is $n_l = n_+ + n_-$. The standard kernel discriminant analysis [9] learns the kernel K and direction $w \in \mathcal{H}_k$ via the optimization problem

$$\begin{aligned} \max_{w,K} \ F(w, K) &= \frac{(w^T(\mu_K^+ - \mu_K^-))^2}{w^T(\Sigma_K^+ + \Sigma_K^- + \lambda I)w} \\ \text{s.t.} \quad K &= \sum_{i=1}^{p} \theta_i K_i, \ \mathbf{1}^T\theta = 1, \ \theta \succeq 0, \end{aligned} \tag{2}$$

where $K_1, \ldots, K_p$ be the given p based kernels, $\theta \succeq 0$ means its elements θ_i are nonnegative, $\lambda > 0$ is a regularization parameter, I is the identity operator in $\mathcal{H}_K$, μ_K^+ and μ_K^- are the sample means

$$\mu_K^+ = \frac{1}{n_+} \sum_{i=1}^{n_+} \varphi_K(x_i), \quad \mu_K^- = \frac{1}{n_-} \sum_{i=n_++1}^{n_l} \varphi_K(x_i),$$

and Σ_K^+ and Σ_K^- are the sample covariances

$$\Sigma_K^+ = \frac{1}{n_+}\sum_{i=1}^{n_+}(\varphi_K(x_i) - \mu_K^+)(\varphi_K(x_i) - \mu_K^+)^T,$$
$$\Sigma_K^- = \frac{1}{n_-}\sum_{i=n_+ +1}^{n_l}(\varphi_K(x_i) - \mu_K^-)(\varphi_K(x_i) - \mu_K^-)^T.$$

Note that (2) is a supervised learning model which neglects the knowledge of the unlabeled data, and hence can not yield the favorable classification. In addition, this model requires all samples to come from the identical distribution, which means that it can not deal with the cross-domain problem. Motivated by this, we propose a domain transfer kernel learning method in the next subsection.

3.2 Domain Transfer Kernel Learning for Discriminant Analysis

Note that if the distributions $\mathcal{P}(x)$ and $\mathcal{Q}(x)$ are completely independent, then the out-of-domain data $\mathcal{D}^{out}$ is useless; if $\mathcal{P}(x)$ and $\mathcal{Q}(x)$ are identical, then the cross-domain problem becomes the standard classification problem. However, in most cases $\mathcal{P}(x)$ and $\mathcal{Q}(x)$ are neither independent nor identical, for which we may use the cross-domain projection vector that is learned from out-of-domain data set $\mathcal{D}^{out}$ and the in-domain data set $\mathcal{D}^{in}$ with MMD formulation. Then, the optimization problem of DTDKL can be formulated as:

$$\max_{w,K} \ \ \mathrm{KLDA}_{K,w}(\mathcal{D}_l) - \beta \mathrm{MMD}_K^2(\mathcal{D}^{out}, \mathcal{D}^{in}), \tag{3}$$

where $\beta \geq 0$ is a parameter to balance the difference of data distributions of two domains and the Fisher discriminant ratio of KLDA for labeled samples. This optimization problem involves two classes of variables. One is the kernel matrix K which represents the adaptive feature space, and the other is the projection direction w for the dimensionality reduction.

By specializing $\mathrm{KLDA}_{K,w}(\mathcal{D}_l)$ and $\mathrm{MMD}_K^2(\mathcal{D}^{out}, \mathcal{D}^{in})$ as $F(w,K)$ and $\|\mu^{in} - \mu^{out}\|$, respectively, (3) becomes

$$\begin{aligned} \max_{w,K} \ \ & F_{\lambda,\beta}(w,K) \\ \text{s.t.} \ \ & K = \sum_{i=1}^{p}\theta_i K_i, \ \mathbf{1}^T\theta = 1, \ \theta \succeq 0 \end{aligned} \tag{4}$$

where

$$F_{\lambda,\beta}(w,K) = \frac{(w^T(\mu_K^+ - \mu_K^-))^2}{w^T(\Sigma_K^+ + \Sigma_K^- + \lambda I)w} - \beta \left\|\mu_K^{in} - \mu_K^{out}\right\|^2. \tag{5}$$

and $\mu_K^+, \mu_K^-, \Sigma_K^+, \Sigma_K^-$ denote the training samples' (in-domain and out-of-domain) means and covariances, respectively. Comparing with the model (2), we see that a new term $-\|\mu^{in} - \mu^{out}\|^2$ is introduced into the objective of (4). This term is a concave function that will bring a concavification effect on the original nonconcave objective of (2). So, the globally optimal solution of the maximization problem (4) can be easier found than that of the problem (2) proposed in [9] .

Note that the last term in $F_{\lambda,\beta}(w, K)$ is independent of w. Hence, the maximization problem (4) can be rewritten as

$$\begin{aligned} &\max_K \ \max_w \ \left\{ \frac{(w^T(\mu_K^+ - \mu_K^-))^2}{w^T(\Sigma_K^+ + \Sigma_K^- + \lambda I)w} \right\} - \beta \left\| \mu_K^{in} - \mu_K^{out} \right\|^2 \\ &\text{s.t.} \quad K = \sum_{i=1}^{p} \theta_i K_i, \ \mathbf{1}^T\theta = 1, \ \theta \succeq 0. \end{aligned} \tag{6}$$

Using the same arguments as in [9], we know that the globally optimal solution of the inner maximization problem

$$w^* = (\Sigma_K^+ + \Sigma_K^- + \lambda I)^{-1}(\mu_K^+ - \mu_K^-). \tag{7}$$

Substituting this into the objective of (8), we obtain that

$$\begin{aligned} &\max_K \ F^*_{\lambda,\beta}(K) \\ &\text{s.t.} \quad K = \sum_{i=1}^{p} \theta_i K_i, \ \mathbf{1}^T\theta = 1, \ \theta \succeq 0 \end{aligned} \tag{8}$$

where

$$\begin{aligned} F^*_{\lambda,\beta}(K) = (\mu_K^+ - \mu_K^-)^T(\Sigma_K^+ + \Sigma_K^- + \lambda I)^{-1}(\mu_K^+ - \mu_K^-) \\ -\beta \left\| \mu_K^{in} - \mu_K^{out} \right\|^2. \end{aligned} \tag{9}$$

On the other hand, from the Representer Theory [12], the optimal discriminative projection in DTDKL is the span of the images of the training points in the feature space. Note that in this method the training set includes both labeled data and unlabeled data due to the MMD formulation. Hence, there exists a vector $\alpha \in \mathbb{R}^n$ such that

$$w^* = \sum_{i=1}^{n} \alpha_i^* \varphi_K(x_i) = U_K \alpha^* \tag{10}$$

where

$$U_K = [\varphi_K(x_1) \ \cdots \ \varphi_K(x_n)] \,.$$

In fact, we can find a closed-form expression of α:

$$\alpha^* = \frac{1}{\lambda}[I - G(\lambda I + GKG)^{-1}GK]a \tag{11}$$

where a is an n-dimensional vector given by

$$a = [1/n_+, \ldots, 1/n_+, -1/n_-, \ldots, -1/n_-, 0, \ldots 0]^T \in \mathbb{R}^n,$$

and the matrix G is defined as

$$G = \begin{pmatrix} \frac{1}{\sqrt{n_+}}(I - \frac{1}{n_+}\mathbf{1}_{n_+}\mathbf{1}_{n_+}^T) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{1}{\sqrt{n_-}}(I - \frac{1}{n_-}\mathbf{1}_{n_-}\mathbf{1}_{n_-}^T) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Because the unlabel data is introduced in this model, the representation of the variables are different from those of [9]. By equations (7), (10) and (11), $F^*_{\lambda,\beta}(K)$ can be written as

$$\begin{aligned} F^*_{\lambda,\beta}(K) &= (\mu_K^+ - \mu_K^-)^T(\Sigma_K^+ + \Sigma_K^- + \lambda I)^{-1}(\mu_K^+ - \mu_K^-) \\ &\quad -\beta b^T K b \\ &= (\mu_K^+ - \mu_K^-)^T w^* - \beta b^T K b \\ &= a^T U_K^T U_K \alpha^* - \beta b^T K b \\ &= \frac{1}{\lambda} a^T K (I - G(\lambda I + GKG)^{-1} GK) a - \beta b^T K b \end{aligned}$$

where $b = (b_1, \ldots, b_n)$ with

$$b_i = \begin{cases} \frac{1}{n^{out}} & \text{if} \quad x_i \in \mathcal{D}^{out}; \\ -\frac{1}{n^{in}} & \text{if} \quad x_i \in \mathcal{D}^{in}. \end{cases} \tag{12}$$

Then, the optimization problem (8) can be reformulated as

$$\begin{aligned} \min_{\theta,t} \quad & -\frac{1}{\lambda}\sum_{i=1}^{p} \theta_i (a^T K_i a - \lambda\beta b^T K_i b) + t \\ \text{s.t.} \quad & a^T KG(\lambda I + GKG)^{-1} GKa \le t, \\ & \mathbf{1}^T\theta = 1, \ \theta \succeq 0. \end{aligned} \tag{13}$$

By the Schur Complement Theorem, we know that

$$a^T KG(\lambda I + GKG)^{-1} GKa \le t \Leftrightarrow \begin{pmatrix} \lambda I + GKG & GKa \\ a^T KG & t \end{pmatrix} \succeq 0.$$

The last two equations show that (13) is equivalent to

$$\begin{aligned} \min_{t \in \mathbb{R}, \theta \in \mathbb{R}^p} \quad & \frac{1}{\lambda}\left(t - \sum_{i=1}^{p} \theta_i (a + \sqrt{\beta} b)^T K_i (a + \sqrt{\beta} b)\right) \\ \text{s.t.} \quad & S(t,\theta) \succeq 0 \\ & \mathbf{1}^T\theta = 1, \quad \theta \succeq 0, \end{aligned} \tag{14}$$

where

$$S(t,\theta) = \begin{pmatrix} \sum_{i=1}^p \theta_i J^T K_i G + \lambda I & \sum_{i=1}^p \theta_i G^T K_i a \\ \sum_{i=1}^p \theta_i a^T K_i G & t \end{pmatrix}.$$

Algorithm 1. Kernel Discriminant Learning in Cross-domain Problem

```
input : A labeled out-of-domain data set $\mathcal{D}^{out}= \{x_i^{out}, y_i^{out}\}$, an unlabeled
        in-domain data set $\mathcal{D}^{in} = \{x_i^{in}\}$ and positive parameters $\lambda$, $\beta$.
output: Labels $Y^{in}$ of the unlabeled data $X^{in}$ in the in-domain.
1. Solve SDP problem in (14) to obtain a kernel matrix $K$
2. Compute the coefficient vector $\alpha$ through (11), then the direction $w$ can be
   obtained from (10)
3. Use the direction $w$ to get new representations $\{x_i^{out'}\}$ and $\{x_i^{in'}\}$ of the original
   data $\{x_i^{out}\}$ and $\{x_i^{in}\}$, respectively.
4. Train a classifier or regressor: $f : x_i^{out'} \to y_i^{out'}$
5. Use the trained classifier or regressor to predict the labels
```

Thus, we convert the nonconvex optimization problem (4) into a convex semidefinite programming problem. Similar to the one obtained by [9], it can be solved by interior-point method softwares such as SeDuMi or SDPT3.

The cost of constructing the basic kernel matrices is $O(n^2d)$, the combining the p basic kernel matrices costs $O(n^2p)$, and computing the gradient and Hessian of the objective is $O(n^3)$. so the total cost per Newton step of interior-point methods which can solve SDP is $O(p^3 + n^2d + n^2p + n^3)$. In the case of $p, d \ll n$, the total cost grows like $O(n^3)$, which is the same as that of SVMs. The SDP guarantees the convergence of the algorithm.

4 Experiment

In this work, we carried out experiments on three real-world data collections from two different domains to evaluate the described algorithms. The performance is compared with MKDL-DA [9], and Semi-supervised kernel discriminant analysis SKDA [13] as well as other transferred dimensionality reduction method, TKDR [2] and MMDE [3].

4.1 Data Sets and Experiment Setup

As shown in Table 1, the data collections consist of Reuters-21578 [14], 20-Newsgroups [15] and SyskillWebert [14]. Amoung them, Reuters-21578 and 20 Newsgroups is the standard used to test web page ratings. The important statistics and pre-processing procedures of these collections are presented below.

Data Sets Description. With a hierarchical structure, SyskillWebert database consists of the HTML source of web pages plus the ratings of a user on those web pages. Four separate subjects are contained in the web pages. Associated with each web page are the HTML source and a user's rating in terms of "hot", "medium" or "cold" [16]. As demonstrated in Table 1, all of the four subjects are involved in our study. "Goat" is reserved as the set of in-domain and the other are used as the out-of-domain data. Compared to the "cold" pages, the total number of pages rated as ”medium” or ”hot” is fewer. Hence, we combine the ”medium” and ”hot” pages together, and change the labels of those pages as ”non-cold” to form a binary classification problem. The learning task is to predict the user's preferences for the given web pages. the Rueters-21578 is another text repository which consists of Reuters news wire articles organized into five top categories, and each category contains various sub-categories. Three categories, ”orgs”, ”people” and ”places”, we remove all the documents of ”USA” in order to make the size of these three categories nearly even [16]. For each category, all of the sub-categories are then organized into two parts, and each part has different distribution and approximately equal size. Therefore, one part can be used for the in-domain and the other is treated as the out-of-domain purpose. According to the method described in [17], three cross-domain learning tasks are generated as listed in Table 1, and the learning objective aims to classify articles into top categories. Similar to Reuters-21578 data, 20-Newsgroups corpus contains 7 top categories and these top categories contain 20 subcategories which have approximately 20,000 newsgroup documents. We select four top categories ”com”, ”rec”, ”talk” and ”sci” in this experiment. Thus, three other cross-domain tasks are formed as listed in Table 1.

Table 1. Summary of Datasets

Data Set		In-domain	Out-of-domain
SyskillWebert		Goat	Bands Sheep Biomedical
Reuters	Orgs vs People Orgs vs Places People vs Places	Documents of some sub categories	Documents of other sub categories
20 News-group	Com vs Rec Rec vs Sci Rec vs Talk	Documents of some sub categories	Documents of other sub categories

Experiment Setup. On one hand, for each in-domain data set employed in the experiment, we further split it into two parts: in-domain data with labels($\mathcal{D}_l$) and the in-domain data without labels($\mathcal{D}_u$). We randomly select 50% data from out-of-domain and in-domain, respectively. The ratio between $|\mathcal{D}_l|$ and $|\mathcal{D}_u|$ is 1:9. All of the in-domain data without labels($\mathcal{D}_u$) are used as the test sets while the training sets consist of the data points with labels from both the in-domain $\mathcal{D}_l$ and out-of-domain ($\mathcal{D}^{out}$). On the other hand, the kernel is a convex combination of 10 Gaussian kernels [10]:

$$K(x,z) = \sum_{i=1}^{10} \theta_i e^{\|x-z\|^2/\sigma_i^2}$$

where θ_i are the weights of the kernels to be determined. The values of σ_i were chosen uniformly over the interval $[10^{-1}, 10^2]$ on the logarithmic scale. The regularization parameter λ in DTDKL and the MMD parameter β was fixed to 10^{-6} and 1, respectively. As a matter of fact, the algorithm is not sensitive to the parameter β for a wide range.

Any ordinary classifier, such as Naïve Bayes, K-nearest, can be used in the dimensionality reduction method. In our experiments, we simply choose the nearest centroid method.

4.2 Experimental Results

For performance evaluation, we use accuracy, which has been widely used as a evaluation metric, we systematically compare the proposed algorithms to some classifiers, including discriminant MKL-DA [9], SKDA [13], as well as TKDR [2], MMDE[3]. All of the results reported below are mean of that running 10 times.

Table 2. Comparison of Performance (*mean ± std %)*

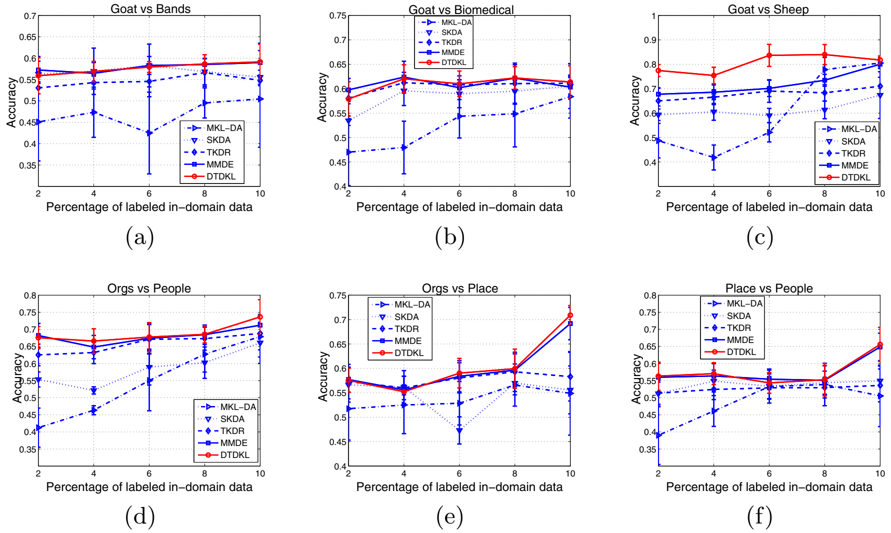
Data Set		MKL-DA	SKDA	TKDR	MMDE	DTDKL
Syskill - Webert	Goat-Bands	50.47 (11.31)	55.56 (1.63)	54.81 (1.24)	57.98 (4.52)	**59.03 (1.13)**
	Goat-Biomedical	58.38 (8.40)	60.54 (8.60)	61.14 (1.13)	60.32 (2.21)	**61.38 (1.30)**
	Goat-Sheep	80.52 (2.25)	67.38 (9.58)	71.05 (3.62)	80.04 (1.33)	**81.75 (1.65)**
Reuters	Orgs-People	67.08 (5.96)	66.05 (6.02)	68.80 (4.81)	71.22 (3.14)	**73.65 (5.05)**
	Orgs-Places	54.90 (8.56)	55.60 (4.91)	58.31 (4.98)	69.19 (3.31)	**70.90 (1.98)**
	People-Places	50.54 (8.98)	54.94 (4.32)	53.60 (4.80)	64.86 (4.01)	**65.60 (4.98)**
20 News-group	Com-Rec	54.49 (9.05)	72.02 (7.42)	72.69 (7.29)	73.05 (3.98)	**73.46 (3.32)**
	Rec-Sci	70.05 (8.78)	73.90 (4.06)	76.90 (5.53)	77.63 (3.29)	**77.68 (4.85)**
	Rec-Talk	70.80 (2.96)	77.35 (5.77)	78.03 (6.64)	78.90 (2.18)	**79.08 (1.52)**

In this section, we use accuracy as the evaluation metric, and compare the proposed algorithms to MKL-DA, SKDA and TKDR. The results show clearly that DTDKL is able to alleviate the influence of different distributions.

Table 2 summarizes the accuracies of MKL-DA, SKDA, TKDR, MMDE, and DTDKL on the three databases with the best results highlighted in bold font. It can be seen that the MAP of the DTDKL methods is consistently higher than the other methods on all of the data sets. Moreover, it is general trend that those problems with higher precision, generally, have a smaller error.

Overall Performance

Our proposed method DTDKL outperforms all the other algorithms in terms of accuracy, demonstrating that DTDKL learns a robust target classifier.And it is also easy to notice that the MMDE and DTDKL performs much better than the other three methods, even TKDR. Moreover, the standard deviation of DTDKL is much smaller, means that it is more stable. For the SyskillWebert collection, compared to DTDKL's rivals, on average it achieves at least 1.23%, 0.24% and 1.05% higher accuracy on "GoatVsBands", "GoatsVsBiomedical" and "GoatVsSheep", respectively. The better performance can be ascribed to transferring the in-domain and out-domain data to a features whose the discriminant



(a) (b) (c)

(d) (e) (f)

Fig. 1. Accuracy vs. different size of $\mathcal{D}_l^{in}$

distance of the data is maximum and the maximum mean discrepancy comes out to be minimum. For the Reuters-21578 data set, the accuracy of DTDKL on average achieve at least 1.6%, 1.7% and 0.7% higher that other approaches on "OrgsVsPeople", "OrgsVsPlaces" and "PeopleVsPlaces" respectively.The similar performance explanation provided to DTDKL method on Reuters-21578 can also applied here. On the 20 News-group data set, the DTDKL methods perform best among the total tasks. Compare DTDKL and MKL-DA, we can see that the MAP of DTDKL is at least 6.6% higher, even nearly 10% higher than MKL-DA, which confirm the positive effect of MMD.

Sensitivity
This study evaluates the sensitivity of varied sizes of labeled in-domain data and conducted on the three collection. The results are demonstrated in Fig.1. It is evident that, as the size of the labeled in-domain data increases, DTDKL performs better than or as equal as its competitors at most case. For example, as shown in Figure 1(c), DTDKL achieves at least 5% higher accuracy than other methods on each size of labeled in-domain data. As a general trend, the accuracy of DTDKL steadily improves when the number of labeled in-domain data increase from 1% to 10%. Consequently, we infer that, better performances can be obtained if more labeled in-domain data are provided.

5 Conclusion

We have proposed a unified dimensionality reduction in cross-domain problems to simultaneously learn a kernel function as well as Fisher discriminant direction by maximizing the Fisher discriminant distance and minimizing the distance of out-of-domain and in-domain. Moreover, we assume that the kernel function in optimal kernel discriminant analysis is a linear combination of multiple base kernels; Thus, it can be efficiently solve by SDP. Experimental result show that DTDKL method outperforms existing dimensionality reduction in cross-domain in three text data sets.

References

1. Mika, S., Ratsch, G., Weston, J., Scholkopf, B., Mullers, K.: Fisher discriminant analysis with kernels. In: NNSP Workshop, pp. 41–48 (1999)
2. Wang, Z., Song, Y., Zhang, C.: Transferred Dimensionality Reduction. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part II. LNCS (LNAI), vol. 5212, pp. 550–565. Springer, Heidelberg (2008)
3. Pan, S., Kwok, J., Yang, Q.: Transfer learning via dimensionality reduction. In: AI, vol. 2, pp. 677–682 (2008)
4. Si, S., Tao, D., Chan, K.: Evolutionary cross-domain discriminative hessian eigenmaps. IEEE Transactions on Image Processing 19(4), 1075–1086 (2010)
5. He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.: Face recognition using laplacianfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(3), 328–340 (2005)
6. Fukunaga, K.: Introduction to statistical pattern recognition. Academic Pr. (1990)
7. Baudat, G., Anouar, F.: Generalized discriminant analysis using a kernel approach. Neural Computation 12(10), 2385–2404 (2000)
8. Lanckriet, G., Cristianini, N., Bartlett, P., Ghaoui, L., Jordan, M.: Learning the kernel matrix with semidefinite programming. The Journal of Machine Learning Research 5, 27–72 (2004)
9. Kim, S.J., Magnani, A., Boyd, S.: Optimal kernel selection in kernel fisher discriminant analysis. In: ICML, pp. 465–472 (2006)
10. Ye, J., Ji, S., Chen, J.: Multi-class discriminant kernel learning via convex programming. The Journal of Machine Learning Research 9, 719–758 (2008)
11. Borgwardt, K., Gretton, A., Rasch, M., Kriegel, H., Schölkopf, B., Smola, A.: Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics 22(14), e49–e57 (2006)
12. Cristianini, N., Shawe-Taylor, J.: Kernel methods for pattern analysis. Cambridge University Press, Cambridge (2004)
13. Cai, D., He, X., Han, J.: Semi-supervised discriminant analysis. In: ICCV, pp. 1–7 (2007)
14. Asuncioin, A., Newman, D.: Uci machine learning repository (2007), `http://www.ics.uci.edu/mlearn/MLRepository.html`
15. Davidov, D., Gabrilovich, E., Markovitch, S.: Parameterized generation of labeled datasets for text categorization based on a hierarchical directory. In: SIGIR, pp. 250–257 (2004)
16. Zhong, E., Fan, W., Peng, J., Zhang, K., Ren, J., Turaga, D., Verscheure, O.: Cross domain distribution adaptation via kernel mapping. In: SIGKDD, pp. 1027–1036 (2009)
17. Dai, W., Yang, Q., Xue, G., Yu, Y.: Boosting for transfer learning. In: ICML, pp. 193–200 (2007)