Cross-Domain Metric Learning Based on Information Theory

Hao Wang[1,2], *__Wei Wang__[2,3], **Chen Zhang**[2], **Fanjiang Xu**[2]
1. State Key Laboratory of Computer Science
2. Science and Technology on Integrated Information System Laboratory
Institute of Software, Chinese Academy of Sciences, Beijing 100190, China
3. Department of Automation, University of Science and Technology of China
weiwangpenny@gmail.com

Abstract

Supervised metric learning plays a substantial role in statistical classification. Conventional metric learning algorithms have limited utility when the training data and testing data are drawn from related but different domains (i.e., source domain and target domain). Although this issue has got some progress in feature-based transfer learning, most of the work in this area suffers from non-trivial optimization and pays little attention to preserving the discriminating information. In this paper, we propose a novel metric learning algorithm to transfer knowledge from the source domain to the target domain in an information-theoretic setting, where a shared Mahalanobis distance across two domains is learnt by combining three goals together: 1) reducing the distribution difference between different domains; 2) preserving the geometry of target domain data; 3) aligning the geometry of source domain data with its label information. Based on this combination, the learnt Mahalanobis distance effectively transfers the discriminating power and propagates standard classifiers across these two domains. More importantly, our proposed method has closed-form solution and can be efficiently optimized. Experiments in two real-world applications demonstrate the effectiveness of our proposed method.

Introduction

Distance metric learning is of fundamental importance in machine learning. Previous research has demonstrated that appropriate distance metrics learnt from labeled training data can greatly improve classification accuracy (Jin, Wang, and Zhou 2009). Depending on whether the geometry information is used, state-of-the-art supervised metric learning methods can be classified into two categories, i.e., globality and locality. Globality metric learning methods aim at keeping all the data points in the same class close together for *compactness* while ensuring those from different classes far apart for *separability* (Davis et al. 2007; Globerson and Roweis 2006; Wang and Jin 2009; Xing et al. 2002). Locality metric learning methods incorporate the geometry of data with the label information to accommodate multimodal data distributions and to further improve classification performance (Weinberger and Saul 2009; Yang et al. 2006). Existing metric learning methods always perform well when there are sufficient labeled training samples. However, in some real-world applications, obtaining the label information of data points drawn from the task-specific domain (i.e., target domain) is extremely expensive or even impossible. One may turn to find labeled data drawn from a related but different domain (i.e., source domain) and apply it as prior knowledge. Apparently, distance metrics learnt only in source domain cannot be directly reused in target domain, although these two domains are closely related. It is because that the significant distribution difference between the data drawn from source and target domains is not explicitly taken into considerations, and this difference will make classifiers trained in source domain invalid in target domain. Therefore, it is important and necessary to reduce the distribution difference between labeled source domain data and unlabeled target domain data in distance metric learning.

Recently, some feature extraction approaches in transfer learning (Caruana 1997; Pan and Yang 2010) have been proposed to address this problem by implicitly exploring a metric (similarity) as a bridge for information transfer from the source domain to the target domain (Geng, Tao, and Xu 2011; Long et al. 2013; Pan, Kowok, and Yang 2008; Pan et al. 2011; Si, Tao, and Geng 2010). These feature extraction methods learn a shared feature representation across domains by 1) reducing the distribution difference, 2) preserving the important properties (e.g., variance or geometry) of data, especially the target domain data. However, most work in this area does not focus on incorporating the geometry with the label information of source domain data to improve the classification performance in target domain. Moreover, these methods formulate a semidefinite programming (SDP) (Boyd and Vandenberghe 2004) or a non-convex optimization problem, resulting in expensive computation.

In this paper, we address the transfer learning problem from the metric learning view and propose a novel algorithm named Cross-Domain Metric Learning (CDML). Specifically, CDML first minimizes the distance between different distributions such that the marginal distributions of target domain and source domain data are close under the learnt distance metric. Second, two Gaussian distributions are con-

*Corresponding author who made main idea and contribution to this work.

structed, one based on the Mahalanobis distance to be learnt and the other based on the geometry of target domain data. By minimizing the relative entropy between these two distributions, the geometry of target domain data is preserved in the learnt distance metric. Third, another two Gaussian distributions are constructed, one based on the Mahalanobis distance to be learnt as well and the other based on the labels and the geometry of source domain data. By minimizing the relative entropy between these two distributions, the learnt distance metric pulls the source domain data in the same class close together, while pushing differently labeled data far apart. Finally, the three terms above are combined into the unified loss function of CDML. This combination effectively transfers the discriminating power gained from the labeled source domain data to the unlabeled target domain data. To the best of our knowledge, our method has made the first attempt to cross-domain metric learning based on relative entropy. We emphasize that CDML has the closed-form solution, leading to efficient optimization.

In summary, the contribution of this paper is two-fold. From the perspective of metric learning, we aim at addressing the challenge of distribution difference. From the perspective of transfer learning, a novel algorithm is proposed to transfer knowledge by finding a shared Mahalanobis distance across domains. The optimal metric can be found efficiently in closed-form. Under this optimal metric, the data distributions are close and points from different classes can be well separated. As a result, we can train standard classifiers in the source domain and reuse them to correctly classify the target domain data. Experimental results in real-world applications verify the effectiveness and efficiency of CDML compared with state-of-the-art metric learning methods and transfer learning methods.

Related Work

Metric Learning

Significant efforts in metric learning have been spent on learning a Mahalanobis distance from labeled training data for classification. Existing Mahalanobis distance learning methods can be classified into two categories, i.e., globality and locality. A natural intention in globality learning is to formulates an SDP for keeping the same labeled points similar (i.e., the distances between them should be small) and differently labeled points dissimilar (i.e., the distances should be larger) (Globerson and Roweis 2006; Xing et al. 2002). Other notable work in globality learning is based on information theory (Davis et al. 2007; Wang and Jin 2009). In particular, Information-Theoretic Metric Learning (ITML) (Davis et al. 2007) formulates the relative entropy as a Bregman optimization problem subject to linear constraints. Information Geometry Metric Learning (IGML) (Wang and Jin 2009) minimizes the Kullback-Leibler (K-L) divergence between two Gaussian distributions and finds the closed-form solution. Locality metric learning methods maximally align the geometry of data with its label information (Weinberger and Saul 2009; Yang et al. 2006) to further improve their performance. However, the supervised algorithms discussed above are limited by the underlying assumption that training data and testing data are drawn from the same distribution.

Transfer Learning

State-of-the-art transfer learning can be organized into instance reweighing (Dai et al. 2007a) and feature extraction. In the feature extraction category, recent work tries to find a subspace shared by both domains, such that the distribution difference is explicitly reduced and the important properties of original data are preserved (Geng, Tao, and Xu 2011; Long et al. 2013; Pan, Kowok, and Yang 2008; Si, Tao, and Geng 2010). In this subspace, classifiers can be propagated between domains. Specifically, Maximum Mean Discrepancy Embedding (MMDE) (Pan, Kowok, and Yang 2008) employs Maximum Mean Discrepancy (MMD) (Gretton et al. 2006) to estimate the distance between different distributions and learns a kernel matrix by preserving the data variance at the same time. Joint Distribution Adaption (JDA) (Long et al. 2013) extends MMD and constructs feature subspace by Principal Component Analysis (PCA) (Jolliffe 1986). Transfer Subspace Learning (TSL) (Si, Tao, and Geng 2010) integrates the Bregman divergence with some dimension reduction algorithms, e.g., PCA and Fisher's linear discriminant analysis (FLDA) (Fisher 1936). However, these methods formulate an SDP or a non-convex optimization, which has high computational complexity and requires iteratively updating parameters. Even worse, the non-convex problems are prone to being trapped in local solutions. In comparison, our metric learning method has efficient closed-form solution and optimally transfers the discriminating power. We would also like to mention that Transfer Component Analysis (TCA) (Pan et al. 2011) is an efficient kernel learning method to extend MMDE. Our work differs from TCA significantly in the proposed optimization. In this paper, an optimal Mahalanobis distance is searched by utilizing the relationship between Gaussian distributions.

Cross-Domain Metric Learning Based on Information Theory

In this section, we present the proposed algorithm named Cross-Domain Metric Learning (CDML) in detail.

Problem Definition

We begin with the problem definition. Table 1 lists the important notations used in this paper.

Definition 1. ***(The Mahalanobis Distance)*** *Denote $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d$, and then the Mahalanobis distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ is calculated as follows:*

$$d_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{A} (\mathbf{x}_i - \mathbf{x}_j), \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{d \times d}$ is positively semi-definite.

In fact, there is a close link between Mahalanobis distance and linear transformation. If we define a linear projection $\mathbf{W}$: $\mathbf{W}^T\mathbf{W} = \mathbf{A}$ which maps $\mathbf{x}_i$ to $\mathbf{W}\mathbf{x}_i$, the Euclidean distance between $\mathbf{W}\mathbf{x}_1$ and $\mathbf{W}\mathbf{x}_2$, i.e., $\|\mathbf{W}\mathbf{x}_1 - \mathbf{W}\mathbf{x}_2\|^2 = (\mathbf{x}_1 - \mathbf{x}_2)^T\mathbf{W}^T\mathbf{W}(\mathbf{x}_1 - \mathbf{x}_2) = (\mathbf{x}_1 - \mathbf{x}_2)^T\mathbf{A}(\mathbf{x}_1 - \mathbf{x}_2)$, is actually the Mahalanobis distance between $\mathbf{x}_1$ and $\mathbf{x}_2$.

Table 1: List of important notations used in this paper.

Notation	Description
$\mathbf{X}_{src} = \{(\mathbf{x}_1^s, y_1^s), ..., (\mathbf{x}_n^s, y_n^s)\}$	Source domain data set
$\mathbf{X}_{tar} = \{\mathbf{x}_1^t, ..., \mathbf{x}_m^t\}$	Target domain data set
$\mathbf{X} = \{\mathbf{x}_1^s, ..., \mathbf{x}_n^s, \mathbf{x}_1^t, ..., \mathbf{x}_m^t\}$	Input data set
$\mathbf{W}$	Linear transformation matrix
$\mathbf{A} = \mathbf{W}^T\mathbf{W}$	Mahalanobis distance matrix
$\mathbf{L}$	The MMD matrix
$\mathbf{K}_{tar} = [\langle \mathbf{W}\mathbf{x}_i^t, \mathbf{W}\mathbf{x}_j^t \rangle]_{m\times m}$	The linear kernel matrix for $\mathbf{W}\mathbf{X}_{tar}$
$\mathbf{K}_T$	The ideal kernel matrix for $\mathbf{X}_{tar}$
$\mathbf{K}_{src} = [\langle \mathbf{W}\mathbf{x}_i^s, \mathbf{W}\mathbf{x}_j^s \rangle]_{n\times n}$	The linear kernel matrix for $\mathbf{W}\mathbf{X}_{src}$
$\mathbf{K}_S$	The ideal kernel matrix for $\mathbf{X}_{src}$

Problem 1. ***(Cross-Domain Metric Learning Based on Information Theory)*** *Let $\mathbf{X}_{tar}$ be a set of m unlabeled testing samples drawn from a target domain: $\mathbf{X}_{tar} = \{\mathbf{x}_1^t, ..., \mathbf{x}_m^t\}$, where $\mathbf{x}_i^t \in \mathbb{R}^d$. Let $\mathbf{X}_{src}$ be a set of n labeled training samples drawn from a related source domain: $\mathbf{X}_{src} = \{(\mathbf{x}_1^s, y_1^s), ..., (\mathbf{x}_n^s, y_n^s)\}$, where $\mathbf{x}_i^s \in \mathbb{R}^d$ and $y_i^s \in \mathcal{Y}^s$ is the class label. We denote $P_t(\mathbf{X}_{tar})$ and $P_s(\mathbf{X}_{src})$ as the marginal probability distributions of $\mathbf{X}_{tar}$ and $\mathbf{X}_{src}$ respectively, $P_t(\mathbf{X}_{tar}) \neq P_s(\mathbf{X}_{src})$. Our task is to learn a shared metric distance $\mathbf{A}$ across domains under which 1) the distribution difference between $P_s(\mathbf{X}_{src})$ and $P_t(\mathbf{X}_{tar})$ is explicitly reduced; 2) the geometry of $\mathbf{X}_{tar}$ is preserved; 3) the points from $\mathbf{X}_{src}$ with the same label are kept similar according to the geometry and others are kept dissimilar.*

Minimizing Distribution Difference

Conventional Mahalanobis distance learning methods performs well in the classification setting based on the assumption that training and testing points are drawn from the same distribution (i.e., $P_s(\mathbf{X}_{src}) = P_t(\mathbf{X}_{tar})$). When such a distance metric $\mathbf{W}_c$ is learnt from $\mathbf{X}_{src}$, it can improve classification accuracy on $\mathbf{X}_{tar}$ using standard classifiers such as KNN and SVM. However, $P_s(\mathbf{X}_{src})$ is usually different from $P_t(\mathbf{X}_{tar})$ since $\mathbf{X}_{src}$ and $\mathbf{X}_{tar}$ are drawn from different but related domains. In this case, $P_s(\mathbf{W}_c\mathbf{X}_{src})$ and $P_t(\mathbf{W}_c\mathbf{X}_{tar})$ are still significantly different and standard classification models trained on $\mathbf{W}_c\mathbf{X}_{src}$ cannot be directly applied on $\mathbf{W}_c\mathbf{X}_{tar}$. Therefore, it is necessary to find a metric $\mathbf{W}$ which can reduce the distance between different distributions. This issue is of particular importance and gains its popularity in transfer learning. Inspired by the work (Long et al. 2013; Pan, Kowok, and Yang 2008), we adopt the criterion *Maximum Mean Discrepancy* (MMD) to measure the distance between $P_s(\mathbf{W}\mathbf{X}_{src})$ and $P_t(\mathbf{W}\mathbf{X}_{tar})$. The empirical estimate of MMD is as follows:

$$\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{W}\mathbf{x}_i^s - \frac{1}{m}\sum_{i=1}^{m} \mathbf{W}\mathbf{x}_i^t \|^2 = tr(\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{A}), \quad (2)$$

where $\mathbf{X} = \{x_1^s, ..., x_n^s, x_1^t, ..., x_m^t\} \in \mathbb{R}^{d\times(n+m)}$, $\mathbf{L} \in \mathbb{R}^{(n+m)\times(n+m)}$ with:

$$\mathbf{L}(i,j) = \begin{cases} \frac{1}{n^2} & \text{if } \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}_{src} \\ \frac{1}{m^2} & \text{if } \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}_{tar} \\ -\frac{1}{nm} & \text{otherwise.} \end{cases} \quad (3)$$

By minimizing Equation (2), $P_s(\mathbf{W}\mathbf{X}_{src})$ and $P_t(\mathbf{W}\mathbf{X}_{tar})$ are close to each other.

Transferring Discriminating Power Based on Information Theory

The metric distance $\mathbf{W}$ learnt by only minimizing the distribution difference may merge all data points together, which is unsuitable for the classification task. To improve classification accuracy, as stated in Problem 1, $\mathbf{W}$ should combine minimizing the distribution difference with 1) preserving the geometry of $\mathbf{X}_{tar}$, 2) maximally aligning the geometry of $\mathbf{X}_{src}$ with its label information. Based on this combination, it is supposed that $P_s(\mathcal{Y}^s|\mathbf{W}\mathbf{X}_{src}) \approx P_t(\mathcal{Y}^t|\mathbf{W}\mathbf{X}_{tar})$. $\mathbf{W}$ optimally transfers discriminating power gained from the source domain to the target domain, that is, the same labeled points are kept close together and the differently labeled points are pushed far apart. In this way, if a classifier is trained on $\mathbf{W}\mathbf{X}_{src}$ and $\mathcal{Y}^s$, it can be reused to correctly classify $\mathbf{W}\mathbf{X}_{tar}$. Note that the combination can perform well because $\mathbf{X}_{tar}$ and $\mathbf{X}_{src}$ share some latent variables.

Geometry Preservation of $\mathbf{X}_{tar}$ Preserving the geometry of unlabeled $\mathbf{X}_{tar}$ is particular useful for transfer learning (Long et al. 2012; Wang and Mahadevan 2011; Pan et al. 2011). We construct a linear kernel $\mathbf{K}_{tar}$ for $\mathbf{W}\mathbf{X}_{tar}$:

$$\mathbf{K}_{tar} = (\mathbf{W}\mathbf{X}_{tar})^T(\mathbf{W}\mathbf{X}_{tar}) = \mathbf{X}_{tar}^T\mathbf{A}\mathbf{X}_{tar}. \quad (4)$$

To introduce the information theory into the space of positive definite matrices, $\mathbf{K}_{tar}$ is related as the covariance matrix of a multivariate Gaussian distribution with zero mean (Wang and Jin 2009):

$$Pr(\mathbf{z}|\mathbf{K}_{tar}) = \frac{1}{(2\pi)^{m/2}|\mathbf{K}_{tar}|^{1/2}} exp(-\mathbf{z}^T\mathbf{K}_{tar}^{-1}\mathbf{z}/2), \quad (5)$$

where $\mathbf{z} \in \mathbb{R}^m$. In the ideal case, an ideal kernel matrix $\mathbf{K}_T$ is expected to give a useful similarity such that the geometry of $\mathbf{X}_{tar}$ is preserved. $\mathbf{K}_T$ is related as the covariance matrix of another multivariate Gaussian distribution:

$$Pr(\mathbf{z}|\mathbf{K}_T) = \frac{1}{(2\pi)^{m/2}|\mathbf{K}_T|^{1/2}} exp(-\mathbf{z}^T\mathbf{K}_T^{-1}\mathbf{z}/2), \quad (6)$$

where $\mathbf{z} \in \mathbb{R}^m$. The distance between $\mathbf{K}_{tar}$ and $\mathbf{K}_T$, denoted as $d(\mathbf{K}_{tar}\|\mathbf{K}_T)$, can be derived by the K-L divergence between the two distributions in Equation (5) and (6):

$$\begin{aligned} d(\mathbf{K}_{tar}\|\mathbf{K}_T) &= KL(Pr(\mathbf{z}|\mathbf{K}_{tar})\|Pr(\mathbf{z}|\mathbf{K}_T)) \\ &= \int Pr(\mathbf{z}|\mathbf{K}_{tar}) log \frac{Pr(\mathbf{z}|\mathbf{K}_{tar}))}{Pr(\mathbf{z}|\mathbf{K}_T)} d\mathbf{z}. \end{aligned} \quad (7)$$

Theorem 1. *The distance between $\mathbf{K}_{tar}$ and $\mathbf{K}_T$ in Equation (7) is equivalent to:*

$$d(\mathbf{K}_{tar}\|\mathbf{K}_T) = \frac{1}{2}(tr(\mathbf{K}_T^{-1}\mathbf{K}_{tar}) - log|\mathbf{K}_{tar}| + log|\mathbf{K}_T| - m). \quad (8)$$

To capture the information of $\mathbf{K}_T$, the optimal $\mathbf{A}$ is searched by minimizing the distance $d(\mathbf{K}_{tar}\|\mathbf{K}_T)$ in Equation (8). Therefore, the geometry of unlabeled $\mathbf{X}_{tar}$ can be preserved in the learnt distance $\mathbf{A}$:

$$\begin{aligned} \mathbf{A} &= arg\min_{\mathbf{A}\succeq 0} d(\mathbf{K}_{tar}\|\mathbf{K}_T) \\ &= arg\min_{\mathbf{A}\succeq 0} tr(\mathbf{K}_T^{-1}\mathbf{X}_{tar}^T\mathbf{A}\mathbf{X}_{tar}) - log|\mathbf{X}_{tar}^T\mathbf{A}\mathbf{X}_{tar}|. \end{aligned} \quad (9)$$

The remaining issue is to define the ideal kernel $\mathbf{K}_T$ for geometry preservation.

1. Constructing a k-nearest neighbor graph: let $\mathbf{G}^t$ denote a directed graph containing a set of nodes $\mathbf{V}^t$ numbered 1 to m and a set of edges $\mathbf{E}^t$. Two nodes i and j are connected by an edge (i.e., $(i,j) \in \mathbf{E}^t$) if x_i^t is one of the k nearest neighbor of x_j^t.

2. Choosing weights: let $\mathbf{M}^t$ refer to the adjacency matrix of $\mathbf{G}^t$, and it is given by:

$$\mathbf{M}^t(i,j) = \begin{cases} exp(-\frac{d_{ij}}{2\sigma^2}) & \text{if } (i,j) \in \mathbf{E}^t \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

where $d_{ij} = \|\mathbf{x}_i^t - \mathbf{x}_j^t\|^2$ and σ is the width.

3. Defining a kernel function $\mathbf{K}_T$ on $\mathbf{G}^t$: specific kernel functions (Kondor and Lafferty 2002; Smola and Kondor 2003) on $\mathbf{G}^t$ induced by the weights can give a useful and more global sense of similarity between instances. Let $\mathbf{D}^t$ be an $m \times m$ diagonal matrix with $\mathbf{D}_{ii}^t = \sum_j \mathbf{M}_{ij}^t$. The Laplacian of $\mathbf{G}^t$ is $\mathbf{L}^t = \mathbf{D}^t - \mathbf{M}^t$, and the Normalized Laplacian is $\tilde{\mathbf{L}}^t = (\mathbf{D}^t)^{-\frac{1}{2}}\mathbf{L}(\mathbf{D}^t)^{-\frac{1}{2}}$. The eigenvalues and eigenvectors of $\tilde{\mathbf{L}}^t$ are denoted as λ_i^t and ϕ_i^t, i.e., $\tilde{\mathbf{L}}^t = \sum_i \lambda_i^t(\phi_i^t)(\phi_i^t)^T$. In this paper, we investigate the diffusion kernel (Kondor and Lafferty 2002) which is proven to be a generalization of Gaussian kernel to graphs:

$$\mathbf{K}_T = \sum_{i=1}^{m} exp(-\sigma_d^2/2\lambda_i^t)(\phi_i^t)(\phi_i^t)^T, \tag{11}$$

where $\mathbf{K}_T \succ 0$ since all the eigenvalues are positive (i.e., $exp(-\sigma_d^2/2\lambda_i^t) > 0$).

Label Information Utilization of $\mathbf{X}_{src}$ A linear kernel $\mathbf{K}_{src}$ is constructed for $\mathbf{WX}_{src}$: $\mathbf{K}_{src} = (\mathbf{WX}_{src})^T(\mathbf{WX}_{src}) = \mathbf{X}_{src}^T\mathbf{AX}_{src}$. Label information is critical for classification tasks and encourages the similarities between two points if and only if they belong to the same class. Geometry preservation is an important component for generalization ability (Weinberger and Saul 2009; Yang et al. 2006). By incorporating these two sources of information, an ideal kernel $\mathbf{K}_S$ is defined for $\mathbf{X}_{src}$ based on two idealizations: 1) similarities between points with different labels will be penalized; 2) similarities between points in the same class will be encouraged according to the neighborhood structure.

1. Constructing a within class graph: let $\mathbf{G}^s$ denote a directed graph which consists of a set of nodes $\mathbf{V}^s$ numbered 1 to n and a set of edges $\mathbf{E}^s$. Two nodes i and j are connected by an edge (i.e., $(i,j) \in \mathbf{E}^s$) if $y_i^s = y_j^s$.

2. Choosing the adjacency matrix $\mathbf{M}^s$ of $\mathbf{G}^s$: $\mathbf{M}^s(i,j) = exp(-\frac{d_{ij}}{2\sigma^2})$ if $(i,j) \in \mathbf{E}^s$, otherwise $\mathbf{M}^s(i,j) = 0$.

3. Defining a diffusion kernel function $\mathbf{K}_S$ on $\mathbf{G}^s$: $\mathbf{K}_S = \sum_{i=1}^{n} exp(-\sigma_d^2/2\lambda_i^s)(\phi_i^s)(\phi_i^s)^T$, where (λ_i^s, ϕ_i^s) are eigenvalues and eigenvectors of the Normalized Laplacian.

4. Minimizing $d(\mathbf{K}_{src}\|\mathbf{K}_S)$: the optimal $\mathbf{A}$ is searched by minimizing the distance $d(\mathbf{K}_{src}\|\mathbf{K}_S)$ derived from Equation (8). Therefore, the learnt distance $\mathbf{A}$ maximally aligns the geometry of $\mathbf{X}_{src}$ with its label information:

$$\mathbf{A} = arg\min_{\mathbf{A}\succeq 0} tr(\mathbf{K}_S^{-1}\mathbf{X}_{src}^T\mathbf{AX}_{src}) - log|\mathbf{X}_{src}^T\mathbf{AX}_{src}|. \tag{12}$$

The Cost Function

CDML aims at searching the optimal distance metric $\mathbf{A}$ by minimizing Equation (2), Equation (9) and Equation (12) simultaneously. This combination effectively transfers the discriminating power gained from the labeled source domain data to the unlabeled target domain data. The overall cost function is as follows:

$$\begin{aligned} \mathbf{A} = arg\min_{\mathbf{A}\succeq 0} & tr(\mathbf{X}(\mathbf{K}+\mu\mathbf{L})\mathbf{X}^T\mathbf{A}) - log|\mathbf{X}_{tar}^T\mathbf{AX}_{tar}| \\ & - log|\mathbf{X}_{src}^T\mathbf{AX}_{src}|, \end{aligned} \tag{13}$$

where $\mu > 0$ is a tradeoff and $\mathbf{K} \succ 0 = \begin{pmatrix} \mathbf{K}_S^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_T^{-1} \end{pmatrix}$.

Proposition 1. *The $(n+m) \times (n+m)$ matrix $\mathbf{L}$ in Equation(2) and Equation (13) is positive semi-definite.*

Proof. For any column vector $\mathbf{z} \in \mathbb{R}^{n+m}$, we have

$$\mathbf{z}^T\mathbf{Lz} = \begin{pmatrix} \mathbf{a} & \mathbf{b} \end{pmatrix} \begin{pmatrix} \mathbf{P} & \mathbf{R} \\ \mathbf{R}^T & \mathbf{Q} \end{pmatrix} \begin{pmatrix} \mathbf{a}^T \\ \mathbf{b}^T \end{pmatrix} \tag{14}$$

where $\mathbf{a} = (z_1, ..., z_n)$, $\mathbf{b} = (z_{n+1}, ..., z_{n+m})$, $\mathbf{P} \in \mathbb{R}^{n\times n}$ with $[\mathbf{P}]_{ij} = 1/n^2$, $\mathbf{Q} \in \mathbb{R}^{m\times m}$ with $[\mathbf{Q}]_{ij} = 1/m^2$ and $\mathbf{R} \in \mathbb{R}^{n\times m}$ with $[\mathbf{R}]_{ij} = -1/nm$.

$\mathbf{z}^T\mathbf{Lz}$ in Equation (14) is equal to:

$$\begin{aligned} & \mathbf{aPa}^T + \mathbf{bQb}^T + 2\mathbf{aRb}^T \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{z_i}{n}\frac{z_j}{n} + \sum_{i=1}^{m}\sum_{j=1}^{m}\frac{z_{n+i}}{m}\frac{z_{n+j}}{m} - 2\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{z_i}{n}\frac{z_{n+j}}{m} \\ &= (\frac{z_1}{n} + ... + \frac{z_n}{n} - \frac{z_{n+1}}{m} - ... - \frac{z_{n+m}}{m})^2 \geq 0 \end{aligned}$$

Therefore, $\mathbf{L} \succeq 0$. The proposition follows. □

Based on Proposition 1, we can obtain the closed-form solution of CDML in the following proposition.

Proposition 2. *The optimal solution to Equation (13) is:*

$$\mathbf{A} = 2(\mathbf{X}(\mathbf{K}+\mu\mathbf{L})\mathbf{X}^T)^{-1} \tag{15}$$

Proof. The derivative of Equation (13) w.r.t. $\mathbf{A}$ is:

$$\mathbf{X}(\mathbf{K}+\mu\mathbf{L})\mathbf{X}^T - 2\mathbf{A}^{-1}. \tag{16}$$

Since $\mathbf{K} \succ 0$ and $\mathbf{L} \succeq 0$, then $(\mathbf{K}+\mu\mathbf{L}) \succ 0$. Proposition 2 now follows by setting the derivative to 0. □

Low Dimensional Projections

The Mahalanobis distance metric $\mathbf{A}$ learnt in CDML is of full rank. If $\mathbf{A}$ has the rank $r < d$, we can represent it in the form: $\mathbf{A} = \mathbf{W}_r^T\mathbf{W}_r$, where $\mathbf{W}_r \in \mathbb{R}^{r\times d}$ projects the original data to an r-dimensional space for dimension reduction. To compute $\mathbf{W}_r$, a straightforward solution is to optimize Equation (13) with a constraint $rank(\mathbf{A}) = r$. However, rank constraints on matrices are not convex (Boyd and Vandenberghe 2004). In this paper, the projection matrix $\mathbf{W}_r$ is computed by a substitute approach (Globerson and Roweis 2006) as follows: 1) eigenvalues and eigenvectors of full-rank $\mathbf{A}$ in Equation (15) are calculated: $\mathbf{A} = \sum_{i=1}^{d}\lambda_i\mathbf{u}_i\mathbf{u}_i^T$, where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d$; 2) $\mathbf{W}_r = diag(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})[\mathbf{u}_1^T; \ldots; \mathbf{u}_r^T]$. The eigen spectrum of $\mathbf{A}$ usually rapidly decays and many eigenvalues are very small, suggesting this solution is close to the optimal one returned by minimizing the rank constrained optimization.

Experiments

In this section, we evaluate the proposed method in two metric learning related applications: 1) face recognition and 2) text classification.

Data Preparation

Face Data Sets FERET (Phillips et al. 2000) and YALE (Belhumeur, Hespanha, and Kriegman 1997) are two public face data sets. FERET data set contains 13,539 face images from 1,565 individuals with different sizes, poses, illuminations and facial expressions. YALE data set has 165 images from 15 individuals with different expressions or configurations. Some example face images are shown in Figure 1. As in the previous work (Si, Tao, and Geng 2010), we construct two cross-domain data sets: 1) *Y vs F*: the source domain set is YALE, and the target domain set consists of 100 individuals randomly selected from FERET. 2) *F vs Y*: the source set contains 100 individuals randomly selected from FERET, and the target set is YALE.

Figure 1: Image examples in (a) FERET data set and (b) YALE data set.

Text Data Sets 20-Newsgroups and Reuters-21578 are two benchmark text data sets widely used for evaluating the transfer learning algorithms (Dai et al. 2007b; Li, Jin, and Long 2012; Pan et al. 2011). 20-Newsgroups consists of nearly 20,000 documents partitioned into 20 different subcategories. The corpus has four top categories and each top category has four subcategories as shown in Table 2. Following the work (Dai et al. 2007b), we construct six cross-domain data sets for binary text classification: *comp vs rec*, *comp vs sci*, *comp vs talk*, *rec vs sci*, *rec vs talk* and *sci vs talk*. Specifically, for each data set (e.g., *comp vs rec*), one top category (i.e., comp) is selected as the positive class and the other category (i.e., rec) is the negative class. Then two subcategories under the positive and the negative classes respectively are selected to form the source domain, the other two subcategories are used to form the target domain.

Table 2: Top categories and their subcategories.

Top Category	Subcategory	Examples
comp	comp.graphics, comp.sys.mac.hardware, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware	3870
rec	rec.autos, rec.motorcycles, rec.sport.baseball, rec.sport.hokey	3968
sci	sci.crypt, sci.electronics, sci.med, sci.space	3945
talk	talk.politics.guns, talk.politics.mideast, talk.politics.misc, talk.religion.misc	3250

Reuters-21578 has three biggest top categories: *orgs*, *people* and *places*. The preprocessed version of Reuters-21578 on the web site (http://www.cse.ust.hk/TL/index.html) is used which contains three cross-domain data sets: *orgs vs people*, *orgs vs place* and *people vs place*.

Baseline Methods

We systematically compare CDML with three state-of-the-art metric learning methods, i.e., Information-Theoretic Metric Learning (ITML) (Davis et al. 2007); Information Geometry Metric Learning (IGML) (Wang and Jin 2009); Large Margin Nearest Neighbor (LMNN) (Weinberger and Saul 2009); and three feature-based transfer learning methods, i.e., Joint Distribution Adaption (JDA) (Long et al. 2013); Semisupervised Transfer Component Analysis (SSTCA) (Pan et al. 2011); Transferred Fisher's Linear Discriminant Analysis (TFLDA) (Si, Tao, and Geng 2010);

For the six comparison methods, the parameters spaces are empirically searched using their own optimal parameter settings and the best results are reported. CDML involves four parameters: σ_d, σ, μ and k. Specifically, we set σ_d by searching the values among $\{0.1, 1, 10\}$, σ among $\{0.1, 1, 10\}$ and μ among $\{0.01, 0.1, 1, 10\}$. The neighborhood size k for CDML is 3. In general, CDML is found to be robust to these parameters. The experiments are carried out on a single machine with Intel Core 2 Quad @ 2.40Ghz and 10 GB of RAM running 64-bit Windows 7.

Experimental Results

Results of Face Recognition In this section, we evaluate the ability of CDML to separate different classes in target domain. For *Y vs F* and *F vs Y*, one random point for each target domain class is selected as the reference data set (Si, Tao, and Geng 2010). The dimensionality of each image is reduced to 100 by PCA. All the methods are trained as a metric learning procedure without the labels of target domain data. At the testing stage, the distance between a target point and every reference point is calculated using the learnt distance metric, then the label of the testing point is predicted as that of the nearest reference point. Since FERET and YALE has different class numbers, JDA is not suitable for this task which requires that source and target domain should share the same class number. TFLDA can find at most $c-1$ meaningful dimensions, where c is the class number of source domain. Figure 2 shows the classification error rates across different dimensions. Some observations can be concluded.

The first general trend is that conventional metric learning algorithms (i.e., ITML, IGML and LMNN) show their limits on these cross-domain data sets. The metrics learnt only from the source domain data fail to separate different classes in target domain. The second general trend is that SSTCA shows good classification performance. SSTCA tries to learn a kernel matrix across domains such that the label dependence is maximized and the manifold structure is preserved. However, CDML consistently provides much higher accuracy than SSTCA. A possible reason is that CDML focuses on keeping the data points in the same class close together while ensuring those from different classes far apart. The third general trend is that although TFLDA works

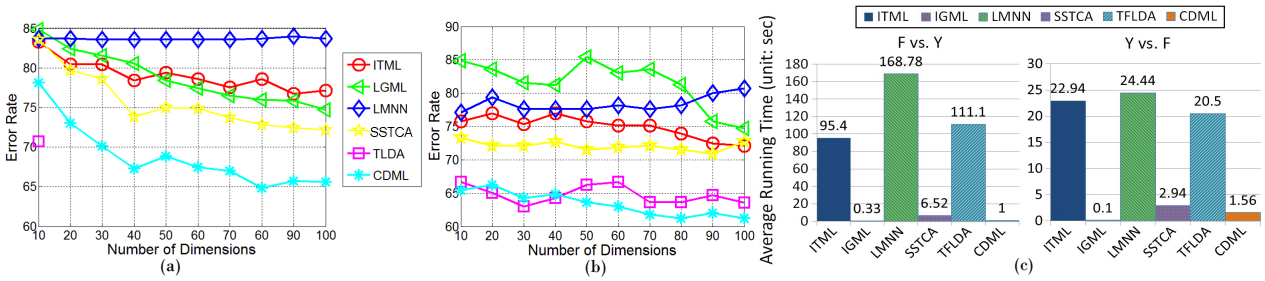

Figure 2: Comparison of ITML, IGML, LMNN, SSTCA, TFLDA and CDML on the face data sets. (a) Classification error rates on Y-F data set. (b) Classification error rates on F-Y data set. (c) Running time comparison.

Table 3: 1-NN classification errors (in percent) of the applied methods.

Method	# Dim	Data Set								
		orgs vs people	*orgs vs place*	*people vs place*	*comp vs rec*	*comp vs sci*	*comp vs talk*	*rec vs sci*	*rec vs talk*	*sci vs talk*
ITML	10	47.43	53.21	45.40	42.91	58.81	59.32	54.85	**43.34**	53.04
	20	58.61	60.02	52.18	43.31	60.37	60.16	52.02	45.87	51.81
	30	52.24	55.99	51.81	42.11	61.80	55.57	53.21	46.92	48.72
IGML	10	52.48	51.39	48.10	47.93	50.32	49.59	50.35	47.94	47.49
	20	52.57	52.92	47.35	46.64	50.04	49.76	50.56	49.13	46.99
	30	52.81	52.73	50.70	46.41	50.29	50.65	52.12	47.72	47.62
LMNN	10	51.32	52.09	49.76	**40.83**	54.52	53.29	51.77	45.39	45.73
	20	51.71	52.87	48.55	**41.86**	54.06	53.55	51.72	**45.73**	45.34
	30	50.45	51.64	48.23	**41.35**	55.17	54.39	51.84	**45.73**	46.18
JDA	10	53.39	56.29	**41.04**	44.62	57.55	48.09	60.98	56.94	52.52
	20	49.59	53.79	**42.80**	48.77	57.81	**41.95**	59.12	56.23	**38.29**
	30	53.73	49.95	**42.62**	47.44	56.81	46.25	60.01	56.11	**41.28**
SSTCA	10	45.45	**46.50**	45.22	49.56	48.37	53.26	50.53	55.60	**45.10**
	20	44.68	46.31	44.29	50.09	47.09	51.91	52.32	49.42	46.18
	30	**44.34**	47.94	45.96	50.67	46.94	47.66	53.34	47.52	45.91
CDML	10	**44.85**	49.27	46.03	47.51	**45.83**	**42.15**	**50.01**	51.19	46.61
	20	**44.62**	**45.64**	44.98	47.18	**45.91**	43.36	**51.20**	50.07	47.18
	30	45.51	**44.87**	45.10	47.41	**46.19**	**45.35**	**49.55**	48.89	45.06

quite well, it can just find at most $c-1$ meaningful dimensions. By contrast, CDML almost achieves the optimal error rate across all the dimensions which illustrates its effective performance in separating different target classes.

To test the efficiency of CDML, we report the average training time in Figure 2(c). ITML, LMNN and TFLDA are computationally expensive since they formulate an alternative optimization problem. Even worse, TFLDA is non-convex and may be trapped in local solutions. Although IGML is fast due to the closed-form solution, it shows high classification error on these cross-domain data sets. We find CDML and SSTCA run quite efficiently, while CDML outperforms SSTCA in terms of classification accuracy.

Results of Text Classification In this section, we evaluate the ability of CDML for text classification and a simple measurement is used: misclassification rate by 1-nearest neighbor classifier (1-NN) without parameters tuning. The unlabeled target instances are compared to the points in the labeled source domain using the learnt distance metric. We compare our proposed CDML with ITML, IGML, LMNN, JDA, SSTCA for this binary task. The classification results across different dimensions are shown in Table 3. Some advantages can be concluded from the results. First, the results of non-transfer metric learning methods are better than that of the transfer algorithms on *comp vs rec* and *rec vs talk*. A possible explanation is that on these two data sets, the distributions of source and target data are not significantly varied. But we would like to mention that the transfer methods always perform well on other cross-domain data sets. Second, JDA provides better results on *people vs place* and *sci vs talk*. The possible explanation is two-fold. 1) Besides reducing the marginal distribution difference, the conditional distribution difference is also exploited in JDA. 2) The common assumption in transferring learning that reducing the difference of marginal distributions will draw close the conditional distributions is not always valid. Third, CDML achieves the minimal error rate on most of the data sets, which illustrates the reliable and effective performance of CDML for domain adaption.

Conclusion

In this paper, we have proposed a novel metric learning algorithm to address transfer learning problem based on information theory. It learns a shared Mahalanobis distance across domains to transfer the discriminating power gained from the source domain to the target domain. Based on the learnt distance, a standard classification model trained only in the source domain can correctly classify the target domain data. Experiments demonstrate the effectiveness of our proposed method. In future work, it is important and promising to explore an online algorithm for cross-domain metric learning and the nonlinear version needs to be investigated.

Acknowledgments

This work is supported by Natural Science Foundation of China (61303164) and Beijing Natural Science Foundation (9144037).

References

Belhumeur, P. N.; Hespanha, J. P.; and Kriegman, D. J. 1997. Eigenfaces versus fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(7):711-720.

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

Caruana, R. 1997. Multitask learning, *Machine Learning* 28(1):41-75.

Dai, W.; Yang, Q.; Xue, G.; and Yu, Y. 2007. Boosting for transfer learning. In *Proceedings of the 24th International Conference on Machine Learning* (ICML), 193-200.

Dai, W.; Xue, G.-R.; Yang, Q.; and Yu, Y. 2007. Coclustering based classification for out-of-domain documents. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD).

Davis, J. V.; Kulis, B.; Jain, P.; Sra, S.; and Dhillon, I. S. 2007. Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning* (ICML), 209-216.

Fisher, R. 1936. The use of multiple measurements in taxonomic problems. Annals of Human Genetics 7(2):179-188.

Geng, B.; Tao, D.; and Xu, C. 2011. DAML: Domain adaptation metric learning. *IEEE Transactions on Image Process* 20(10): 2980-2989.

Globerson, A., and Roweis S. 2006. Metric learning by collapsing classes. In *Proceedings of the 20th Annual Conference on Advances in Neural Information Processing Systems* (NIPS), 451-458.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Scholkopf, B.; and Smola, A. J. 2006. A kernel method for the two-sample problem. In *Proceedings of the 16th Annual Conference on Advances in Neural Information Processing Systems* (NIPS).

Jin, R.; Wang, S.; and Zhou, Y. 2009. Regularized distance metric learning:theory and algorithm. In *Proceedings of the 23rd Annual Conference on Advances in Neural Information Processing Systems* (NIPS), 862-870.

Jolliffe, I. 1986. *Principal Component Analysis*. Springer-Verlag.

Kondor, R. S., and Lafferty, J. 2002. Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the 19th International Conference on Machine Learning* (ICML), 315-322.

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).

Long, M.; Wang, J.; Ding, G.; Sun, J.; and Yu, P. S. 2013. Transfer Feature Learning with Joint Distribution Adaptation. In *Proceedings of the 14th IEEE International Conference on Computer Vision* (ICCV).

Long, M.; Wang, J.; Ding, G.; Shen, D.; and Yang, Q. 2012. Transfer learning with graph co-regularization. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence* (AAAI).

Pan, S. J.; Kwok, J. T.; and Yang, Q. 2008. Transfer learning via dimensionality reduction. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence* (AAAI).

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.

Phillips, J. P.; Moon, H.; Rizvi, S. A.; and Rauss, P. J. 2000. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(10):1090-1104.

Si, S.; Tao, D.; and Geng, B. 2010. Bregman divergence-based regularization for transfer subspace learning. *IEEE Transactions on Knowledge and Data Engineering* 22(7):929-942.

Smola, A., and Kondor, R. 2003. Kernels and regularization on graphs. In *Proceedings of the 16th Annual Conference on Learning Theory* (COLT), 144-158.

Wang, C., and Mahadevan, S. 2011. Heterogeneous domain adaptation using manifold alignment. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence* (AAAI).

Wang, S., and Jin, R. 2009. An information geometry approach for distance metric learning. In *Proceedings of the 12nd International Conference on Artificial Intelligence and Statistics* (AISTATS), 591-598.

Weinberger, K. Q.; Sha, F.; and Saul, L. K. 2004. Learning a kernel matrix for nonlinear dimensionality reduction. In *Proceedings of the 21th International Conference on Machine Learning* (ICML), 839-846.

Weinberger, K. Q., and Saul, L. K. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research* 10:207-244.

Xing, E. P.; Ng, A. Y.; Jordan, M. I.; and Russell, S. J. 2002. Distance metric learning, with application to clustering with side-information. In *Proceedings of the 16th Annual Conference on Advances in Neural Information Processing Systems* (NIPS), 505-512.

Yang, L.; Jin, R.; Sukthankar, R.; and Liu, Y. 2006. An efficient algorithm for local distance metric learning. In *Proceedings of the 21st AAAI Conference on Artificial Intelligence* (AAAI), 543-548.