# General Heterogeneous Transfer Distance Metric Learning via Knowledge Fragments Transfer

**Yong Luo[†], Yonggang Wen[†], Tongliang Liu[‡], Dacheng Tao[‡]**
[†]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[‡]UBTech Sydney AI Institute and SIT, FEIT, The University of Sydney, Australia
yluo180@gmail.com, ygwen@ntu.edu.sg, tongliang.liu@sydney.edu.au, dacheng.tao@sydney.edu.au

## Abstract

Transfer learning aims to improve the performance of target learning task by leveraging information (or transferring knowledge) from other related tasks. Recently, transfer distance metric learning (TDML) has attracted lots of interests, but most of these methods assume that feature representations for the source and target learning tasks are the same. Hence, they are not suitable for the applications, in which the data are from heterogeneous domains (feature spaces, modalities and even semantics). Although some existing heterogeneous transfer learning (HTL) approaches is able to handle such domains, they lack flexibility in real-world applications, and the learned transformations are often restricted to be linear. We therefore develop a general and flexible heterogeneous TDML (HTDML) framework based on the knowledge fragment transfer strategy. In the proposed HTDML, any (linear or nonlinear) distance metric learning algorithms can be employed to learn the source metric beforehand. Then a set of knowledge fragments are extracted from the pre-learned source metric to help target metric learning. In addition, either linear or nonlinear distance metric can be learned for the target domain. Extensive experiments on both scene classification and object recognition demonstrate superiority of the proposed method.

## 1 Introduction

We often encounter the label or side information (such as the similar/dissimilar constraints) deficiency problem in the machine learning and pattern recognition applications due to the high labeling cost (labor-intensive and expensive). Transfer learning [Pan and Yang, 2010; Luo *et al.*, 2014; Hu *et al.*, 2015; Isele *et al.*, 2016] is able to mitigate this problem in the target learning task or domain by leveraging information from other related source tasks or domains [Liu *et al.*, 2017]. A typical example is the unconstrained human gait recognition, which is to identify a person's manner of walking from a distance in the wild conditions [Tao *et al.*, 2007b]. We can only label a few gait images (or image sequences) in a new scenario but we may have large amounts of labeled ones in some other scenarios. The data distributions of two different scenarios can be very different due to the varying lighting and background. Hence, the recognition model trained in a label-rich scenario may perform bad in the label-scarce one, and transfer learning is helpful in this case. Some other examples include the sentiment classification [Pan and Yang, 2010], and image super-resolution [Dai *et al.*, 2015].

Recently, transfer distance metric learning (TDML) [Luo *et al.*, 2014; Hu *et al.*, 2015] has attracted an increasing attention, because it is crucial to learn a reliable distance metric [Tao *et al.*, 2007a] to reveal the data relationships in diverse research areas, ranging from clustering and classification to kernel machines and ranking [Kulis, 2012; Lim and Lanckriet, 2014]. Traditional TDML algorithms usually assume the source and target domain share the same feature representation. This assumption may not be valid in practice. For example, the document representations of different languages vary in multilingual document categorization since the utilized vocabularies are different [Luo *et al.*, 2016]. Some recent works on image annotation and retrieval suggest utilizing source features to guide learning a better representation for target features [Qi *et al.*, 2012; Dai *et al.*, 2015]. The feature spaces of source and target domains can be quite different and there is sometimes semantic gap between them (e.g. text and visual features) [Xu *et al.*, 2014; 2015].

There exist some heterogeneous transfer learning (HTL) approaches [Wang and Mahadevan, 2011; Zhang and Yeung, 2011; Qi *et al.*, 2012] that are able to manage heterogeneous representations. These approaches often transform the heterogeneous features into a common subspace, so that the difference between heterogeneous domains is reduced. Most of the HTL methods are not specially designed for distance metric learning (DML), but we can derive a metric from the learned transformation for each domain. Although effective in some cases, the current HTL approaches exhibit two main defects: 1) the source and target transformations are learned together. Consequently, they are not feasible when original source domain data are not available, and only source metric is provided. Besides, learning both the source and target transformations may significantly increase the complexity of the algorithms when the number of samples in the source domain is large; 2) the transformations are restricted to be linear

and thus the performance may be unsatisfactory in many visual analysis-based applications, since the structure of data distribution is nonlinear for most types of the visual features.

Inspired by the knowledge fragment transfer strategy [Vapnik and Izmailov, 2015], we develop a general heterogeneous TDML (HTDML) framework to overcome these defects. In particular, the proposed HTDML first learns the source distance metric by applying existing linear or nonlinear metric learning algorithms on the given labeled source data. Then we extract some knowledge fragments from the learned metric for transfer. In this paper, we assume there are abundant unlabeled samples that have feature representations in both of the source and target domains. By simultaneously minimizing the empirical losses w.r.t. the metric in the target domain, and enforcing the metric to agree with the knowledge fragments on the unlabeled samples, we learn an improved target metric by making use of the additional source information contained in the fragments.

The main advantages of the proposed HTDML are: 1) the source knowledge fragments can be learned offline, we do not have to reuse the original source domain data. Hence, the algorithm can be used in the applications where source domain data are invisible. Besides, any (linear or nonlinear) metric learning algorithms can be adopted to learn the source knowledge fragments. Thus, the method is general, flexible, and easy-to-use; 2) nonlinear metric can be learned for the target domain by incorporating some nonlinear learning technique, such as gradient boosting regression tree (GBRT) [Kedem *et al.*, 2012]. Hence the proposed method can be widely adopted in many applications, especially the challenging visual-analytic based ones. We conduct experiments on two popular applications: scene classification and object recognition. In addition to the Euclidean (EU) and single domain DML baselines, we further compare with several representative heterogeneous transfer learning approaches [Wang and Mahadevan, 2011; Zhang and Yeung, 2011; Qi *et al.*, 2012; Dai *et al.*, 2015]. The results validate the superiority of the proposed HTDML.

# 2 Heterogeneous Transfer Distance Metric Learning

Problem setting: we suppose the training set with (weakly supervised) side information for the target domain is given by $\mathcal{D}_M^L = \{\mathbf{x}_{Mi}^1, \mathbf{x}_{Mi}^2, y_{Mi}\}_{i=1}^{N_M}$, where $\mathbf{x}_{Mi}^1, \mathbf{x}_{Mi}^2 \in \mathbb{R}^{d_M}$, and $y_{Mi} = \pm 1$ indicates $\mathbf{x}_{Mi}^1$ and $\mathbf{x}_{Mi}^2$ are similar/dissimilar to each other. In the target domain, we have only a few samples with side information, and thus DML may perform poorly. Therefore, we assume there exists a relevant source domain with the training set $\mathcal{D}_S^L = \{\mathbf{x}_{Si}^1, \mathbf{x}_{Si}^2, y_{Si}\}_{i=1}^{N_S}$, where $\mathbf{x}_{Si}^1, \mathbf{x}_{Si}^2 \in \mathbb{R}^{d_S}$ belong to different feature space from the target domain samples. In the source domain, either the features are stronger than the target domain [Dai *et al.*, 2015], or the samples (with side information) are abundant, i.e., $N_S \gg N_M$. To enable knowledge transfer, we also assume there are large amounts of unlabeled data that have representations in both the source and target domains, i.e., $\mathcal{D}^U = \{(\mathbf{x}_{Sn}^U, \mathbf{x}_{Mn}^U)\}_{n=1}^{N^U}$, and such data are usually easy to collect in practice [Qi *et al.*, 2012]. Our goal is to learn an appropriate distance metric $A_M$ for the target domain.

## 2.1 Problem Formulation

We propose a general framework for learning distance metric $A_M$ in the target domain by making use of the information from both target and source domain, as well as the unlabeled data. The framework is motivated by the privileged information transfer strategy presented in [Vapnik and Izmailov, 2015], where the knowledge in a privileged space $\tilde{\mathcal{X}}$ is represented as a set of functions $\tilde{\kappa}(\tilde{\mathbf{p}}_c, \tilde{\mathbf{x}}), c = 1, 2, \ldots, r$. Here, $\tilde{\mathbf{p}}_c$ is called fundamental element, which is a vector from the space $\tilde{\mathcal{X}}$, and $\tilde{\kappa}(\tilde{\mathbf{p}}_c, \tilde{\mathbf{x}})$ is called fragment of knowledge with the kernel function $\tilde{\kappa}$. If we choose the quadratic kernel function, i.e., $\tilde{\kappa}(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) = \langle \tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j \rangle^2$, the fundamental elements can be found exactly by solving an eigenvalue problem [Vapnik and Izmailov, 2015]. To transfer knowledge from the privileged space to the original decision space $\mathcal{X}$, some functions $\{\phi_c(\mathbf{x})\}_{c=1}^r$ are found in space $\mathcal{X}$ to approximate the knowledge fragments $\{\tilde{\kappa}(\tilde{\mathbf{p}}_c, \tilde{\mathbf{x}})\}_{c=1}^r$. Here, $\mathbf{x}$ and $\tilde{\mathbf{x}}$ are the representations of a given sample in the original decision space and privileged space respectively.

In [Vapnik and Izmailov, 2015], both the finding of fundamental elements and knowledge transfer are under the theme of support vector machines (SVM), where sufficient data with class labels and corresponding privilege information are provided for training. Our setting is much more challenging in that: 1) we only have weakly supervised side information for limited data; 2) the weakly labeled data in the target domain are scarce and are usually different from the source domain. Consequently, the method presented in [Vapnik and Izmailov, 2015] is not appropriate for DML, and cannot be used in the heterogeneous transfer setting. To tackle this problem, we propose the following heterogeneous transfer distance metric learning (HTDML) framework.

The framework is based on a generalized notion of the Mahalanobis distance [Kulis, 2012]. In the literature of distance metric learning (DML), most methods focus on learn the Mahalanobis distance, which is often denoted as

$$dst_A(\mathbf{x}_i^1, \mathbf{x}_i^2) = (\mathbf{x}_i^1 - \mathbf{x}_i^2)^T A (\mathbf{x}_i^1 - \mathbf{x}_i^2), \tag{1}$$

Here, $A$ is the metric, which is a positive semi-definite matrix and can be factorized as $A = UU^T$. By applying some simple algebraic manipulations, we have $dst_A(\mathbf{x}_i^1, \mathbf{x}_i^2) = \|U\mathbf{x}_i^1 - U\mathbf{x}_i^2\|_2^2$. In order to take the structure of data distribution into consideration, we propose to conduct DML in the feature space determined by a mapping $\psi$, i.e., $dst_A(\mathbf{x}_i^1, \mathbf{x}_i^2) = \|U\psi(\mathbf{x}_i^1) - U\psi(\mathbf{x}_i^2)\|_2^2$. Then the distance can be further denoted as

$$dst_\phi(\mathbf{x}_i^1, \mathbf{x}_i^2) = \|\phi(\mathbf{x}_i^1) - \phi(\mathbf{x}_i^2)\|_2^2, \tag{2}$$

Here, $\phi(\cdot) = U\psi(\cdot)$ is an integrated mapping function, which is unspecified and can be either linear or nonlinear. Then the learning of the target metric $A_M$ is reformulated as learning the mapping $\phi_M$, and the general formulation of the proposed HTDML for learning $\phi_M$ is given by

$$\arg\min_{\phi_M} \epsilon(\phi_M) = E(\phi_M) + \gamma R(\{\phi_{Mc}(\cdot)\}, \{f_{Sc}(\cdot)\}), \tag{3}$$

where $E(\phi_M) = \frac{1}{N_M}\sum_i L(\phi_M; \mathbf{x}^1_{Mi}, \mathbf{x}^2_{Mi}, y_{Mi})$ is the empirical loss w.r.t. $\phi_M$ in the target domain. We choose $L(\phi_M; \mathbf{x}^1_{Mi}, \mathbf{x}^2_{Mi}, y_{Mi}) = g(y_{Mi}[1 - dst_{\phi_M}(\mathbf{x}^1_{Mi}, \mathbf{x}^2_{Mi})])$ and adopt the hinge loss for $g$, i.e., $g(z) = \max(0, b - z)$. Here, $b$ is set to be zero, and the distance between any pair of samples is given by (2). The regularization term $R(\{\phi_{Mc}(\cdot)\}, \{f_{Sc}(\cdot)\})$ is to enforce knowledge transfer from the source domains to the target domain, where $\phi_{Mc}$ is the $c$'th coordinate of the vector-valued mapping function $\phi_M$ and $f_{Sc}(\cdot)$ is the $c$'th fragment of knowledge in source domain. The knowledge transfer is performed by using the mapping functions $\{\phi_{Mc}(\cdot)\}$ in the target domain to approximate the fragments of knowledge in the source domains $\{f_{Sc}(\cdot)\}$. In this way, the knowledge in the source domain is incorporated to learn the mapping function $\phi_M$, i.e., the distance metric in the target domain. The trade-off hyper-parameter $\gamma \geq 0$.

The knowledge fragments in the source domain can be found in various ways. If classification labels are available, we can train SVM classifiers and use the obtained support vectors as the fundamental elements, and then construct the knowledge fragments using a pre-defined kernel. However, in DML, we are often only provided with side (weakly supervised) information, i.e., the similarity/dissimilarity between two samples $\mathbf{x}_i$ and $\mathbf{x}_j$. Therefore, we first learn the metric for source domain using some existing DML algorithm. The output of a metric learning algorithm can be a distance metric $A_S$ [Davis *et al.*, 2007] or feature mapping $\phi_S$ [Kedem *et al.*, 2012]. For the distance metric, we decompose it as $A_S = P_S P_S^T$, and the columns $\{\mathbf{p}_{Sc}\}_{c=1}^r$ of the matrix $P_S$ are adopted as the fundamental elements. Then the source knowledge fragment is given by $f_{Sc}(\cdot) = \kappa_S(\mathbf{p}_{Sc}, \cdot)$, where $\kappa_S$ is a pre-defined kernel in the source domain. For the feature mapping $\phi_S$, the knowledge fragment is directly obtained as $f_{Sc}(\cdot) = \phi_{Sc}(\cdot)$, where $\phi_{Sc}$ is the $c$'th coordinate of $\phi_S$.

Given the pre-trained source knowledge fragment $f_{Sc}(\mathbf{x}^U_{Sn})$, we propose to minimize the divergence between $\phi_{Mc}(\mathbf{x}^U_{Mn})$ and $f_{Sc}(\mathbf{x}^U_{Sn})$, where $\mathbf{x}^U_{Sn}$, $\mathbf{x}^U_{Mn}$ are representations in the source and target domain respectively for a given unlabeled sample. To this end, the regularization term in (3) can be defined as follows,

$$\begin{aligned} &R(\{\phi_{Mc}(\cdot)\}, \{f_{Sc}(\cdot)\}) \\ &= \frac{1}{N^U}\sum_{n=1}^{N^U}\left(\sum_{c=1}^{r}\mathrm{Div}\left(\phi_{Mc}(\mathbf{x}^U_{Mn}), f_{Sc}(\mathbf{x}^U_{Sn})\right)\right), \end{aligned} \tag{4}$$

where $r$ is the number of fundamental elements, and $\mathrm{Div}(\cdot,\cdot)$ is a divergence measure, which can be the absolute difference, and least squares error.

In this paper, we adopt the absolute difference to suppress the effect of outliers. This leads to the following compact regularization term:

$$R(\Phi_M, F_S) = |\Phi_M - F_S|, \tag{5}$$

where

$$\Phi_M = \begin{bmatrix} \phi_{M1}(\mathbf{x}^U_{M1}) & \cdots & \phi_{M1}(\mathbf{x}^U_{MN^U}) \\ \vdots & \ddots & \vdots \\ \phi_{Mr}(\mathbf{x}^U_{M1}) & \cdots & \phi_{Mr}(\mathbf{x}^U_{MN^U}) \end{bmatrix}$$

is the mapped matrix of the unlabeled data in the target domain, and $F_S$ is a knowledge fragment matrix represented by the unlabeled data in the source domain, i.e.,

$$F_S = \begin{bmatrix} f_{S1}(\mathbf{x}^U_{S1}) & \cdots & f_{S1}(\mathbf{x}^U_{SN^U}) \\ \vdots & \ddots & \vdots \\ f_{Sr}(\mathbf{x}^U_{S1}) & \cdots & f_{Sr}(\mathbf{x}^U_{SN^U}) \end{bmatrix}$$

with each $f_{Sc}(\mathbf{x}^U_{Sn}) = \kappa_S(\mathbf{p}_{Sc}, \mathbf{x}^U_{Sn})$ or $f_{Sc}(\mathbf{x}^U_{Sn}) = \phi_{Sc}(\mathbf{x}^U_{Sn})$. Here, $|A| = \sum_i\sum_j |A_{ij}|$ is the sum of all the elements' absolute values of a matrix $A$. By substituting (5) into (3), we obtain the following specific optimization problem for HTDML:

$$\begin{aligned} &\arg\min_{\phi_M} \epsilon(\phi_M) \\ &= \frac{1}{N_M}\sum_i g\left(y_{Mi}[1 - \|\phi_M(\mathbf{x}^1_{Mi}) - \phi_M(\mathbf{x}^2_{Mi})\|_2^2]\right) \\ &\quad + \gamma|\Phi_M - F_S|. \end{aligned} \tag{6}$$

In the following, we first assume $\phi_M = U_M \in \mathbb{R}^{d_M \times r}$ is a linear transformation, and then extend it to the nonlinear case.

## 2.2 Linear Formulation and Optimization

When we choose $\phi_M = U_M$ as a linear transformation, the problem (6) can be reformulated as

$$\begin{aligned} &\arg\min_{U_M} \epsilon(U_M) \\ &= \frac{1}{N_M}\sum_i g\left(y_{Mi}[1 - \|U_M^T(\mathbf{x}^1_{Mi} - \mathbf{x}^2_{Mi})\|_2^2]\right) \\ &\quad + \gamma|U_M^T X_M^U - F_S|, \\ &\text{s.t. } U_M \succeq 0, \end{aligned} \tag{7}$$

where $X_M^U = [\mathbf{x}^U_{M1}, \mathbf{x}^U_{M2}, \ldots, \mathbf{x}^U_{MN^U}] \in \mathbb{R}^{d_M \times N^U}$ is the data matrix of the unlabeled samples in the target domain. The constraint $U_M \succeq 0$ means that each element of $U_M$ is non-negative. This constraint not only narrows the hypothesis space for $U_M$, but also makes the results easy to inspect and interpret.

For notation simplicity, we set $\delta_{Mi} = \mathbf{x}^1_{Mi} - \mathbf{x}^2_{Mi}$, so that $\|U_M(\mathbf{x}^1_{Mi} - \mathbf{x}^2_{Mi})\|_2^2 = \delta_{Mi}^T U_M U_M^T \delta_{Mi}$, and the optimization problem becomes

$$\arg\min_{U_M} \epsilon(U_M) = E(U_M) + \Omega(U_M), \text{ s.t. } U_M \succeq 0, \tag{8}$$

where $E(U_M) = \frac{1}{N_M}\sum_{i=1}^{N_M} g(y_{Mi}[1 - \delta_{Mi}^T U_M U_M^T \delta_{Mi}])$ and $\Omega(U_M) = \gamma|U_M^T X_M^U - F_S|$. We propose to solve the problem (8) efficiently by utilizing the projected gradient method (PGM) presented in [Lin, 2007]. Because the terms in both $E(U_M)$ and $\Omega(U_M)$ are non-differentiable, we first smooth it according to [Nesterov, 2005]. Then the gradient of the smoothed $\epsilon(U_M)$ is

$$\frac{\partial \epsilon^\sigma(U_M)}{\partial U_M} = \frac{1}{N_M}\sum_{i=1}^{N_M}\left(2y_{Mi}\nu_{Mi}(\delta_{Mi}\delta_{Mi}^T)U_M\right) + \gamma X_M^U Q_M^T. \tag{9}$$

where

$$\nu_{Mi} = \text{median}\left\{\frac{-y_{Mi}(1-\delta_{Mi}^T U_M U_M^T \delta_{Mi})}{\sigma\|\delta_{Mi}\|_\infty}, 0, 1\right\}. \quad (10)$$

and $Q_M \in \mathbb{R}^{r\times N^U}$ is a matrix with the entry $q_{M,cn} = \text{median}\left\{\frac{\mathbf{u}_{Mc}^T \mathbf{x}_{Mn}^U - f_{S,cn}}{\sigma}, -1, 1\right\}$. Here, $\mathbf{u}_{Mc}$ is the $c$'th column of $U_M$ and $f_{S,cn}$ is the $(c,n)$'th element of $F_S$; $\sigma$ is the smooth parameter, which is set as $0.5$ in this paper. Due to limited page length, we omit the detailed derivation. Finally, based on the obtained gradient, we apply the improved PGM presented in [Lin, 2007] to minimize the smoothed primal $\epsilon^\sigma(U_M)$, i.e.,

$$U_M^{t+1} = \pi[U_M^t - \mu_t \nabla\epsilon^\sigma(U_M^t)], \quad (11)$$

where the operator $\pi[x]$ projects all the negative entries of $x$ to zero, and $\mu_t$ is the step size that must satisfy the following condition:

$$\epsilon^\sigma(U_M^{t+1}) - \epsilon^\sigma(U_M^t) \le \rho\nabla\epsilon^\sigma(U_M^t)^T(U_M^{t+1} - U_M^t), \quad (12)$$

where the parameter $\rho$ is chosen to be $0.01$ following [Lin, 2007]. The step size can be determined using the Algorithm 1 cited from [Lin, 2007] (Algorithm 4 therein), and the convergence of the algorithm is guaranteed according to [Lin, 2007]. The stopping criterion we utilized here is $|\epsilon^\sigma(U_M^{t+1}) - \epsilon^\sigma(U_M^t)|/|\epsilon^\sigma(U_M^{t+1}) - \epsilon^\sigma(U_M^0)| < \varepsilon$), where the initialization $U_M^0$ is set as a random matrix.

### 2.3 Nonlinear Extension

When we allow the mapping $\phi_M$ to be nonlinear, the problem (6) can be rewritten as

$$\begin{aligned}&\arg\min_{\phi_M} \epsilon(\phi_M)\\ &= \frac{1}{N_M}\sum_{i=1}^{N_M} g\left(y_{Mi}[1 - \|\phi_M(\mathbf{x}_{Mi}^1) - \phi_M(\mathbf{x}_{Mi}^2)\|_2^2]\right)\\ &+ \gamma\sum_{n=1}^{N^U}|\phi_M(\mathbf{x}_{Mn}^U) - \mathbf{f}_{Sn}|,\end{aligned} \quad (13)$$

where $\mathbf{f}_{Sn}$ is the $n$'th column of the matrix $F_S$. To find an appropriate nonlinear form for $\phi_M$, we assume it is a gradient boosting function given by $\phi_M = \phi_M^0 + \alpha\sum_{t=1}^T \hbar_{Mt}$, where $\phi_M^0$ is an initialization, and $\hbar_{Mt}$ is a regression tree, together with a learning rate $\alpha$ [Kedem *et al.*, 2012]. The solution can be obtained by iteratively adding regression trees $\hbar_{Mt}$ to minimize the objective $\epsilon(\phi_M)$ in a greedy way [Friedman, 2001].

In the following, we summarize the procedure of finding the (approximately) optimal tree in each iteration. For notation simplicity, we omit the subscripts $S$ and $M$. In iteration $t$, the (approximately) optimal tree $\hbar_t^*$ is found by selecting a tree from the set of all regression trees $\mathcal{T}^p$ to approximate the negative gradient of $\epsilon(\phi_{t-1})$ w.r.t. $\phi_{t-1}$, which is the mapping learned at the previous iteration. Here, $p$ is the depth of the trees. Similar to the linear formulation, we smooth the non-differentiable terms to calculate the gradients. Then the tree $\hbar_t^*(\cdot)$ is learned by approximating it with the negative gradient $neg_t(\cdot)$ over each training sample, i.e.,

$$\begin{aligned}\hbar_t^*(\cdot) = &\arg\min_{\hbar\in\mathcal{T}^p}\sum_{i=1}^N\left(\hbar(\mathbf{x}_i) - neg_t(\mathbf{x}_i)\right)^2\\ &+\sum_{n=1}^{N^U}\left(\hbar(\mathbf{x}_n^U) - neg_t(\mathbf{x}_n^U)\right)^2,\end{aligned} \quad (14)$$

where $neg_t(\mathbf{x}_i) = -\frac{\partial\epsilon(\phi_{t-1})}{\partial\phi_{t-1}(\mathbf{x}_i)}$. The tree is greedily learned by pGBRT [Tyree *et al.*, 2011]. The problem (13) is non-convex w.r.t. $\phi_M$, so we initialize $\phi_M$ as $\phi_M^0 = U_M^*$, which is the optimal transformation learned by our linear formulation. This makes the extension to be a nonlinear refinement of the linear formulation.

The time complexity of the proposed algorithm for our linear formulation is $O(T_2T_1rd_M(N_M+N^U))$, where $T_1$ is the number of checks that are needed to find the step size, and $T_2$ is the number of iterations for reaching the stop criterion. Suppose the number of trees utilized in the pGBRT algorithm is $\Gamma$, then the time complexity for our nonlinear formulation is $O((T_2T_1r + \Gamma\log(N^B))d_M(N_M+N^U))$, where $N^B \ll (N_M+N^U)$. The complexity is independent on the number samples $N_S$ and feature dimension $d_S$ in the source domain, and linear w.r.t. all the number $r$, $d_M$, $N_M$ and $N^U$. Thus, the proposed algorithm is quite efficient as long as $N^U$ is not very large.

## 3 Experiments

In this section, we evaluate the effectiveness of the proposed HTDML algorithm on both scene classification and object recognition. Prior to these evaluations, we present our experimental settings.

### 3.1 Experimental Setup

The comparison methods are listed as below:

- **EU:** directly computing the Euclidean distance between the normalized representations of different samples in the target domain.
- **LMNN [Weinberger *et al.*, 2005]:** learning the distance metric for the target domain using the large margin nearest neighbor algorithm presented in [Weinberger *et al.*, 2005]. The number of attracted target neighbors is chosen from 1 to 10.
- **ITML [Davis *et al.*, 2007]:** learning the distance metric for the target domain using the information-theoretic metric learning algorithm presented in [Davis *et al.*, 2007]. The trade-off hyper-parameter is tuned over the set $\{10^i|i=-5,-4,\ldots,3,4\}$.
- **MTDA [Zhang and Yeung, 2011]:** a heterogeneous multi-task learning algorithm by extending linear discriminant analysis to handle multiple heterogeneous domains. The hyper-parameter of intermediate dimensionality is set as a fixed value since the model is not very sensitive to it according to [Zhang and Yeung, 2011].

- **DAMA [Wang and Mahadevan, 2011]:** a heterogeneous domain adaptation algorithm by aligning the manifolds of different domains using the class labels. The hyper-parameter is determined according to the strategy presented in [Wang and Mahadevan, 2011].
- **DT [Qi *et al.*, 2012]:** a heterogeneous distance function transfer algorithm by leveraging large amounts of corresponding data between the source and target domain. The similarity of two target domain samples is determined according to the similarities of their corresponding source domain samples. The candidate sets for the two balancing hyper-parameters are both $\{10^i | i = -5, -4, \ldots, 4\}$.
- **MI [Dai *et al.*, 2015]:** a recently proposed metric imitation algorithm. The knowledge transfer is performed by manifold structure approximation between the source and target domains. The hyper-parameters are determined according to [Dai *et al.*, 2015].
- **HTDML:** the proposed heterogeneous transfer distance metric learning algorithm. In the linear formulation, LMNN [Weinberger *et al.*, 2005] is adopted to find the fundamental elements in the source domain. In the nonlinear extension, we employ GB-LMNN [Kedem *et al.*, 2012] to learn the source knowledge fragments. Because GBRT is adopted to learn the mapping in the target domain, we call the proposed nonlinear extension **GB-HTDML**. The hyper-parameter $\gamma$ is optimized over the set $\{10^i | i = -5, -4, \ldots, 3, 4\}$.

The single domain distance metric learning (DML) algorithms (LMNN and ITML) only utilize the limited label or side information in each domain, and do not make use of any additional information from other domains. The heterogeneous transfer learning (HTL) approaches, MTDA and DAMA, mainly utilize the label or side information in the source and target domain to build a connection between them. DAMA also leverage large amounts of unlabeled data to preserve the topology in each domain. They do not aim to learn distance metric, so we derive the metric as $A = UU^T$ after learning the transformation matrix $U \in \mathbb{R}^{d \times r}$ for the target domain. DT and MI are also HTL methods, but they focus on metric learning and perform knowledge transfer by utilizing the unlabeled correspondence information between the source and target domain. The chosen of an optimal dimensionality $r$ of the mapped subspace is still an open problem, and we do not study it in this paper. To this end, for MTDA, DAMA, and the proposed HTDML, we perform comparisons on a set of varied $r$.

In all the following experiments, each feature space is regarded as a domain. The task in the target domain is to perform multi-class classification [Liu and Tao, 2016], where the $k$-nearest neighbor classifier is adopted. Parameter determination is still an open issue in heterogenous transfer learning due to the limited labeled samples in the target domain. Consequently, if unspecified, the hyper-parameters are tuned in the range mentioned above and the best results of different compared methods are reported. Both the classification accuracy and macroF1 [Sokolova and Lapalme, 2009] score are utilized as evaluation criteria. The side information in terms of pairwise similarity constraints are obtained according to whether two labeled training samples belong to the same class or not. The remained training data that have representations in both domains are used as unlabeled data. Ten random choices of the labeled instances or sample pairs are used, and the mean values with standard deviations are reported.

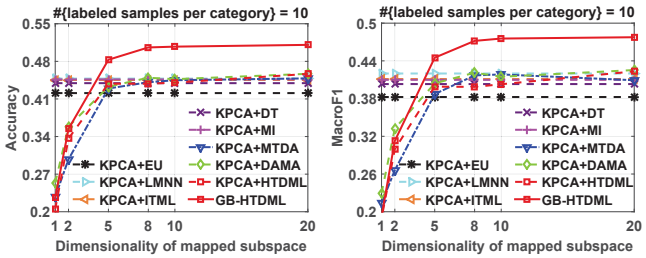


Figure 1: Classification accuracies and macroF1 scores of the nonlinear methods vs. dimensionality of the mapped subspace on the Scene-15 dataset.

## 3.2 Scene Categorization

The dataset used in scene categorization is the Scene-15 [Lazebnik *et al.*, 2006], which contains 4585 images belonging to 15 natural scene categories. We randomly split the image set into a training and test set of equal size. We choose the "expensive" CNN feature [Chatfield *et al.*, 2014] as the source domain, and the "cheap" GIST feature [Oliva and Torralba, 2001] as the target domain. The used features are provided by [Dai *et al.*, 2015], where the feature dimensions of CNN and GIST are 4096 and 20 respectively. We select 10 labeled instances for each category in both the source and target domain to see how much the "expensive" feature can help the metric learning of "cheap" feature [Dai *et al.*, 2015]. In the following, we first select 2000 unlabeled cross-domain correspondences, and then investigate the performance of varying number of correspondences.

For all the compared methods except the proposed GB-HTDML, we preprocess the features by kernel PCA (KPCA) to take the nonlinear structure of the data distribution into consideration. In the proposed linear HTDML with the KPCA preprocess, we adopt the linear kernel (for the preprocessed data) in the source domain. The result dimensions are 2000 and 20 for CNN and GIST respectively. We also compared the different methods without the KPCA preprocess, and found that the performance of all methods are worse than that of using the preprocess. This indicates that the distribution structures of the utilized visual features are indeed nonlinear, and KPCA is able to exploit such nonlinearity to some extent. We do not show the results without preprocess here due to the limited page length.

The classification results of using the preprocess are shown in Fig. 1. We can see from the results that: 1) the HTL approach DAMA is slightly better than the single domain DML algorithms. This is because KPCA helps to build the domain connection by exploiting the nonlinearity, and some source information is transferred to help learn the transformation in the target domain; 2) DT and MI, as well as MTDA

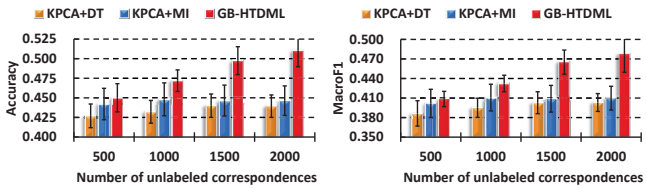


Figure 2: Classification accuracies and macroF1 scores vs. number of unlabeled correspondences on the Scene-15 dataset.

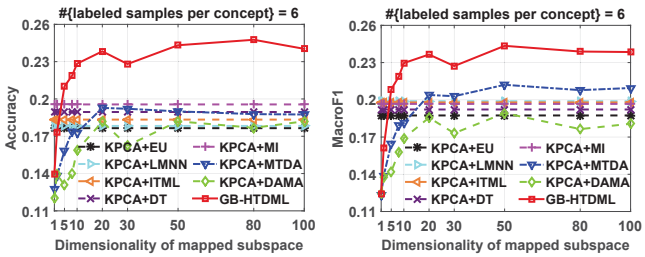


Figure 3: Classification accuracies and macroF1 scores vs. dimensionality of the mapped subspace on the NUS animal subset.

are only comparable to LMNN and ITML. The performance of the proposed HTDML with the KPCA preprocess is only a bit higher than LMNN. This indicates that such a simple preprocess can only bring limited benefits to the transfer approaches, and is not always helpful; 3) the proposed GB-HTDML outperforms all other approaches significantly. This demonstrates that the nonlinearity in the data is successfully captured by the developed nonlinear algorithm. In particular, we obtain significant $11.8\%$ and $12.3\%$ relative improvements over the competitive DAMA in terms of accuracy and macroF1 respectively.

All DT, MI and the proposed HTDML make use of the unlabeled corresponding data between the source and target domain. In this set of experiments, we investigate how the number of correspondences affect their performance. We vary the number from $500$ to $2000$, and the results are shown in Fig. 2. It can be seen from the results that: 1) for DT and MI, there are only small improvements when the number of correspondences increases. Performance of DT and MI reach their peaks around $1500$ and $1000$ respectively. Whereas for the proposed GB-HTDML, the performance improves significantly with an increasing number of correspondences. This demonstrates that the KPCA preprocess cannot effectively capture nonlinearity of the data distribution, which is properly discovered by our GB-HTDML; 2) the improvements of GB-HTDML decrease when the amounts of corresponding data are more than $1500$. This is because there is redundancy in the set of knowledge fragments, and the amount of new knowledge brought by including more correspondences becomes small when the size of the set is large enough.

### 3.3 Object Recognition

We further verify the proposed method in object recognition on a natural image dataset NUS-WIDE (NUS) [Chua *et al.*, 2009]. The dataset contains $269,648$ images, and we conduct experiments on a subset that consists of $16,519$ images belonging to $12$ animal concepts. Half of the images are used for training and the rest for test. In this dataset, we choose the $1000$-D tag feature as source domain, and the $500$-D bag of SIFT [Lowe, 2004] visual words (BOVW) as the target domain. The number of labeled instances for each concept is $6$ in both domains to see how much the easily interpretable text feature can guide the metric learning of visual feature, which is often harder to interpret [Qi *et al.*, 2012]. The number of unlabeled correspondences is $5000$. Original features are used in the proposed GB-HTDML. For other methods, the dimensions of the tag and BOVW representations after the KPCA preprocess are both $300$.

This set of experiments is more challenging than the scene categorization since there is semantic gap between the text and visual features [Qi *et al.*, 2012]. From the results shown in Fig. 3, we can see that: 1) the improvements of the single domain DML algorithms and DT over the EU baseline are not that large as in scene categorization. This indicates that the effectiveness of the KPCA preprocess drops in this application; 2) MI is superior to them and this may benefit from the manifold structure exploited in the source domain; 3) DAMA fail on this dataset because of the semantic gap. MTDA outperforms DAMA since the former learns an additional hidden layer. The transfer is conducted between higher level patterns and thus the semantic gap is reduced; 4) the proposed GB-HTDML outperforms all other approaches significantly when the dimensionality of the mapped subspace is more than $10$. This further verifies the superiority of the proposed method.

## 4 Conclusion

In this paper, we present a general heterogeneous transfer distance metric learning framework. The framework extracts a set of knowledge fragments from the source domain to help the metric learning in the target domain. Any existing distance metric learning (DML) algorithms can be adopted to learn the source knowledge fragments in an offline manner, and either linear or nonlinear metric can be learned for the target domain. Hence the proposed framework is general, flexible, and easy-to-use.

From the experimental evaluation on two popular applications, we mainly conclude that: 1) the performance of most of the current heterogeneous transfer learning (HTL) or metric transfer (imitation) approaches are unsatisfactory in the applications where data lie in a highly nonlinear feature space, or there is a semantic gap between the source and target domain. The KPCA preprocess can sometimes be helpful, but not always take effect; 2) by appropriately exploring the nonlinearity in both the source and target domains, we can obtain significant improvements over the KPCA counterpart.

## Acknowledgments

This work is supported by Singapore MOE Tier 2 (ARC42/13), IMDA Grant GDCR01, NRF via GBIC, administered by BCA, and Australian Research Council Projects FT-130101457, DP-140102164, LP-150100671.

## References


[Chatfield *et al.*, 2014] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[Chua *et al.*, 2009] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS-WIDE: a real-world web image database from national university of singapore. In *CIVR*, 2009.

[Dai *et al.*, 2015] Dengxin Dai, Till Kroeger, Radu Timofte, and Luc Van Gool. Metric imitation by manifold transfer for efficient vision applications. In *IEEE CVPR*, pages 3527–3536, 2015.

[Davis *et al.*, 2007] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216, 2007.

[Friedman, 2001] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

[Hu *et al.*, 2015] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Deep transfer metric learning. In *IEEE CVPR*, pages 325–333, 2015.

[Isele *et al.*, 2016] David Isele, Mohammad Rostami, and Eric Eaton. Using task features for zero-shot knowledge transfer in lifelong learning. In *IJCAI*, pages 1620–1626, 2016.

[Kedem *et al.*, 2012] Dor Kedem, Stephen Tyree, Fei Sha, Gert R Lanckriet, and Kilian Q Weinberger. Non-linear metric learning. In *NIPS*, pages 2573–2581, 2012.

[Kulis, 2012] Brian Kulis. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2012.

[Lazebnik *et al.*, 2006] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE CVPR*, pages 2169–2178, 2006.

[Lim and Lanckriet, 2014] Daryl Lim and Gert Lanckriet. Efficient learning of mahalanobis metrics for ranking. In *ICML*, pages 1980–1988, 2014.

[Lin, 2007] Chih-Jen Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19(10):2756–2779, 2007.

[Liu and Tao, 2016] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE TPAMI*, 38(3):447–461, 2016.

[Liu *et al.*, 2017] Tongliang Liu, Dacheng Tao, Mingli Song, and Stephen J Maybank. Algorithm-dependent generalization bounds for multi-task learning. *IEEE TPAMI*, 39(2):227–241, 2017.

[Lowe, 2004] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[Luo *et al.*, 2014] Yong Luo, Tongliang Liu, Dacheng Tao, and Chao Xu. Decomposition-based transfer distance metric learning for image classification. *IEEE TIP*, 23(9):3789–3801, 2014.

[Luo *et al.*, 2016] Yong Luo, Yonggang Wen, and Dacheng Tao. On combining side information and unlabeled data for heterogeneous multi-task metric learning. In *IJCAI*, pages 1809–1815, 2016.

[Nesterov, 2005] Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

[Oliva and Torralba, 2001] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[Pan and Yang, 2010] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE TKDE*, 22(10):1345–1359, 2010.

[Qi *et al.*, 2012] Guo-Jun Qi, Charu C Aggarwal, and Thomas S Huang. Transfer learning of distance metrics by cross-domain metric sampling across heterogeneous spaces. In *SDM*, pages 528–539, 2012.

[Sokolova and Lapalme, 2009] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427–437, 2009.

[Tao *et al.*, 2007a] Dacheng Tao, Xuelong Li, Xindong Wu, Weiming Hu, and Stephen J Maybank. Supervised tensor learning. *KIS*, 13(1):1–42, 2007.

[Tao *et al.*, 2007b] Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J Maybank. General tensor discriminant analysis and gabor features for gait recognition. *IEEE TPAMI*, 29(10):1700–1715, 2007.

[Tyree *et al.*, 2011] Stephen Tyree, Kilian Q Weinberger, Kunal Agrawal, and Jennifer Paykin. Parallel boosted regression trees for web search ranking. In *WWW*, pages 387–396, 2011.

[Vapnik and Izmailov, 2015] Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: Similarity control and knowledge transfer. *JMLR*, 16:2023–2049, 2015.

[Wang and Mahadevan, 2011] Chang Wang and Sridhar Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *IJCAI*, pages 1541–1546, 2011.

[Weinberger *et al.*, 2005] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, pages 1473–1480, 2005.

[Xu *et al.*, 2014] Chang Xu, Dacheng Tao, and Chao Xu. Large-margin multi-view information bottleneck. *IEEE TPAMI*, 36(8):1559–1572, 2014.

[Xu *et al.*, 2015] Chang Xu, Dacheng Tao, and Chao Xu. Multi-view intact space learning. *IEEE TPAMI*, 37(12):2531–2544, 2015.

[Zhang and Yeung, 2011] Yu Zhang and Dit-Yan Yeung. Multi-task learning in heterogeneous feature spaces. In *AAAI*, pages 574–579, 2011.