Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)

On Combining Side Information and Unlabeled Data for Heterogeneous Multi-Task Metric Learning

Yong Luo[†], **Yonggang Wen**[†], **Dacheng Tao**[‡]
[†]SCSE, Nanyang Technological University, Singapore
[‡]QCIS and FEIT, University of Technology Sydney, Australia
yluo180@gmail.com, ygwen@ntu.edu.sg, dacheng.tao@uts.edu.au

Abstract

Distance metric learning (DML) is critical for a wide variety of machine learning algorithms and pattern recognition applications. Transfer metric learning (TML) leverages the side information (e.g., similar/dissimilar constraints over pairs of samples) from related domains to help the target metric learning (with limited information). Current TML tools usually assume that different domains exploit the same feature representation, and thus are not applicable to tasks where data are drawn from heterogeneous domains. Heterogeneous transfer learning approaches handle heterogeneous domains by usually learning feature transformations across different domains. The learned transformation can be used to derive a metric, but these approaches are mostly limited by their capability of only handling two domains. This motivates the proposed heterogeneous multi-task metric learning (HMTML) framework for handling multiple domains by combining side information and unlabeled data. Specifically, HMTML learns the metrics for all different domains simultaneously by maximizing their high-order correlation (parameterized by feature covariance of unlabeled data) in a common subspace, which is induced by the transformations derived from the metrics. Extensive experiments on both multi-language text categorization and multi-view social image annotation demonstrate the effectiveness of the proposed method.

1 Introduction

Distance metric learning (DML) aims to find an appropriate distance or similarity measure between data. It plays a crucial role in diverse research areas, ranging from the simple k-nearest neighbor (kNN) classification, k-means clustering, to the sophisticated kernel machines (such as support vector machine, or SVM for brief) [Xu *et al.*, 2013] and learning to rank [McFee and Lanckriet, 2010]. It is therefore essential to learn a robust distance metric to reveal the data relationships. To achieve this goal, we need a large amount of side information [Xing *et al.*, 2002] such as the constraints that indicate whether a pair of samples is similar or not.

Recently, some transfer metric learning (TML) [Zha *et al.*, 2009; Zhang and Yeung, 2012] methods were proposed for DML when the side information is scarce in the domain of interest (target domain), while we have abundant side information in certain related, but different source domains [Ammar *et al.*, 2015; Luo *et al.*, 2014]. Traditional DML algorithms usually fail in this scenario because the data distributions between the source and target domain may be quite different, and TML [Zha *et al.*, 2009; Zhang and Yeung, 2012] tries to reduce the impact of such difference and utilize the labeled information from the source domains to help the target metric learning. Specifically, multi-task metric learning (MTML) [Zhang and Yeung, 2012] assume the side information for each of the source and target domains is limited [Goetschalckx *et al.*, 2015; Luo *et al.*, 2013; 2016], and the objective is to improve the metric learning of all domains simultaneously.

One major limitation of most existing TML algorithms is that they assume samples of the related domains are of the same feature dimensionality or lie in the same feature space. This assumption may be not valid for many applications. A typical example is the cross-lingual document classification, where the feature representations of the documents written in different languages vary since the utilized vocabularies are different. Moreover, in multi-view natural image classification and multimedia retrieval, the instances in different domains are often represented in different types of features (such as local SIFT [Lowe, 2004] and global wavelet texture) or have different modalities (such as image, audio and text).

To manage heterogeneous representations, many heterogeneous transfer learning [Shi *et al.*, 2010; Wang and Mahadevan, 2011; Zhou *et al.*, 2014] approaches have been proposed in the literature. A frequently utilized strategy in these methods is to transform the heterogeneous features into a common subspace, where the difference between heterogeneous domains is reduced [Zhou *et al.*, 2014]. The learned transformation for each domain can be used to derive a metric. Although effective in some cases, most of them are limited for only two domains (one source domain and one target domain). However, we usually have more than two domains in many real-world applications. For example, five languages are used in the news articles of the Reuters multilingual collection, and different kinds of local, global, as well as biologically inspired features are popular utilized in visual analysis-based tasks such as image annotation.

To this end, we develop a novel heterogeneous multi-task metric learning (HMTML) framework that handles an arbitrary number of domains by combining side information and unlabeled data. In this paper, we assume there are abundant unlabeled samples that have feature representations in all domains. In particular, HMTML learns the metrics for all different domains in a single optimization problem by minimizing empirical losses w.r.t. the metric for each domain. At the same time, we transform different representations of the given unlabeled samples into a common subspace using the feature transformations derived from the metrics. Because the different representations are modelling the same instance, they should be close to each other in the subspace. By maximizing the high-order covariance between the transformed data representations, we find a shared subspace for all domains and thus all of their side information can be further incorporated to learn this shared subspace of maximum reliability. Hence a more reliable metric is obtained for each domain since the (Mahanobias) metric learning is equivalent to learn a subspace under certain optimization criterion [Kulis, 2012]. Intuitively, the common subspace provides a bridge for side information transfer. In this way, different domains help each other in metric learning, so the learned metrics are more reliable than the results of learning them separately, especially for those domains that have limited side information.

There exist a few approaches [Wang and Mahadevan, 2011; Zhang and Yeung, 2011] that could learn transformations and derive metrics for more than two domains. However, in these approaches, only the statistics (correlation information) between each representation and the shared representation [Zhang and Yeung, 2011], or pairs of representations [Wang and Mahadevan, 2011] is explored, while high-order statistics that can only be obtained by simultaneously examining all domains is ignored. Besides, these approaches mainly focus on utilizing the side information and thus may fail given insufficient side information. Our method is superior to these methods in that we aim to directly maximize the correlation between all domains by analyzing their high-order feature covariance tensor, which is calculated using large amounts of unlabeled data. Much more correlation information can thus be encoded in the learned transformations and also metrics, and hopefully better performance can be achieved. We perform experiments on two popular applications: multi-language text categorization and multi-view social image annotation. In addition to the Euclidean (EU) and single domain regularized DML (RDML) baselines, we further compare the proposed method with two representative heterogeneous transfer learning approaches [Wang and Mahadevan, 2011; Zhang and Yeung, 2011] for multiple domains. The results validate the effectiveness of the proposed HMTML.

2 Heterogeneous Multi-task Metric Learning

In contrast to DAMA [Wang and Mahadevan, 2011] and MTDA [Zhang and Yeung, 2011], which learn linear transformation for each domain by only considering the pairwise

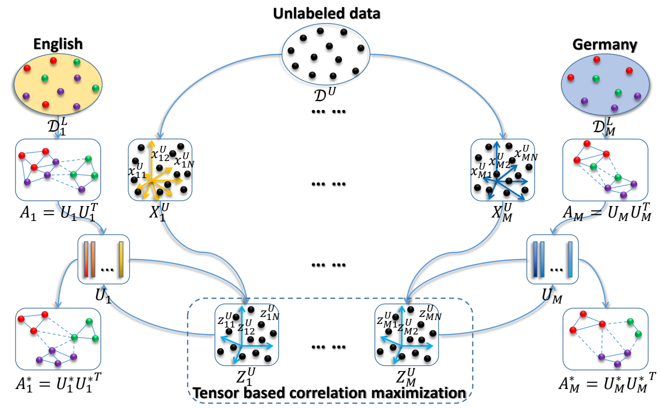

Figure 1: System diagram of the proposed heterogeneous multi-task metric learning (see text for details).

domain correlations, we propose tensor based heterogeneous MTML (HMTML) to learn transformations for metric learning by exploiting the high order tensor correlation between all domains. The diagram of the proposed HMTML is shown in Figure 1. Taking the multilingual text classification as an example, we assume that limited side information (in the form of paired sample similarity) is provided for each of the M heterogeneous domains, such as "English", "Italian", and "Germany". For the m'th domain, we minimize the empirical losses w.r.t. the metric A_m on the labeled data $\mathcal{D}_m^L$. Since the side information is scarce for each domain, learning the different metrics independently may be unreliable. To enable information being shared across all domains so that they can help each other in metric learning, we assume that we are also given abundant unlabeled samples that are represented in all M domains, i.e., $\{\mathbf{x}_{mn}^U\}_{n=1}^{N^U}, m = 1, 2, \ldots, M$. Then we decompose the metric A_m as $A_m = U_m U_m^T$, and using U_m to transform the original heterogeneous representations into a common space as $\{\mathbf{z}_{mn}^U\}_{n=1}^{N^U}, m = 1, 2, \ldots, M$. Finally, by maximizing tensor based high-order covariance between all transformed representations, we learn improved U_m^* by utilizing additional information from other domains, and so more reliable metric $A_m^* = U_m^* (U_m^*)^T$ is obtained. The technical details are given below, and we start by briefing the used notations and concepts of multilinear algebra in this paper.

2.1 Notations

Let $\mathcal{A}$ be an M-order tensor of size $I_1 \times I_2 \times \ldots \times I_M$, and U be a $J_m \times I_m$ matrix. The m-mode product of $\mathcal{A}$ and U is then denoted as $\mathcal{B} = \mathcal{A} \times_m U$, which is an $I_1 \times \ldots \times I_{m-1} \times J_m \times I_{m+1} \ldots \times I_M$ tensor with the element

$$\begin{aligned} &\mathcal{B}(i_1, \ldots, i_{m-1}, j_m, i_{m+1}, \ldots, i_M) \\ &= \sum_{i_m=1}^{I_m} \mathcal{A}(i_1, i_2, \ldots, i_M) U(j_m, i_m). \end{aligned} \tag{1}$$

The product of $\mathcal{A}$ and a sequence of matrices $\{U_m \in \mathbb{R}^{J_m \times I_m}\}_{m=1}^M$ is a $J_1 \times J_2 \times \ldots \times J_M$ tensor denoted by

$$\mathcal{B} = \mathcal{A} \times_1 U_1 \times_2 U_2 \ldots \times_M U_M. \tag{2}$$

The mode-m matricization of $\mathcal{A}$ is denoted as an $I_m \times (I_1 \ldots I_{m-1} I_{m+1} \ldots I_M)$ matrix $A_{(m)}$, which is obtained by mapping the fibers associated with the m'th dimension of $\mathcal{A}$ as the rows of $A_{(m)}$, and aligning the corresponding fibers of all the other dimensions as the columns. Here, the columns can be ordered in any way. The m-mode multiplication $\mathcal{B} = \mathcal{A} \times_m U$ can be manipulated as matrix multiplication by storing the tensors in metricized form, i.e., $B_{(p)} = UA_{(p)}$. Let $\mathbf{u}$ be an I_m-vector, the contracted m-mode product of $\mathcal{A}$ and $\mathbf{u}$ is denoted as $\mathcal{B} = \mathcal{A} \bar{\times}_m \mathbf{u}$, which is an $I_1 \times \ldots \times I_{m-1} \times I_{m+1} \ldots \times I_M$ tensor of order $M-1$, and the entries are calculated by:

$$\begin{aligned} &\mathcal{B}(i_1, \ldots, i_{m-1}, i_{m+1}, \ldots, i_M) \\ &= \sum_{i_m=1}^{I_m} \mathcal{A}(i_1, i_2, \ldots, i_M)\mathbf{u}(i_m). \end{aligned} \tag{3}$$

Finally, the Frobenius norm of the tensor $\mathcal{A}$ is given by

$$\|\mathcal{A}\|_F^2 = \langle \mathcal{A}, \mathcal{A} \rangle = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \ldots \sum_{i_M=1}^{I_M} \mathcal{A}(i_1, i_2, \ldots, i_M)^2. \tag{4}$$

2.2 Problem Formulation

Given M heterogeneous domains, we suppose the training set with side information for the m'th domain is given by $\mathcal{D}_m^L = \{(\mathbf{x}_{mi}, \mathbf{x}_{mj}, y_{mij})\}_{i,j=1}^{N_m}$, where $\mathbf{x}_{mi}, \mathbf{x}_{mj} \in \mathbb{R}^{d_m}$ and $y_{mij} = \pm 1$ indicates $\mathbf{x}_{mi}$ and $\mathbf{x}_{mj}$ are similar/dissimilar to each other. The number of training samples N_m is very small for each domain, so we assume there are large amounts of unlabeled data that have representations in all domains $\mathcal{D}^U = \{(\mathbf{x}_{1n}^U, \mathbf{x}_{2n}^U, \ldots, \mathbf{x}_{Mn}^U)\}_{n=1}^{N^U}$, and these data are usually easy to collect in practice [Qi *et al.*, 2012]. Then the general formulation of the proposed HMTML for learning the metrics $\{A_m\}_{m=1}^M$ is given by

$$\begin{aligned} &\mathop{\arg\min}_{\{A_m\}_{m=1}^M} F(\{A_m\}) = \sum_{m=1}^M \Psi(A_m) + \gamma R(A_1, \ldots, A_M), \\ &\qquad \text{s.t. } A_m \succeq 0, m = 1, 2, \ldots, M, \end{aligned} \tag{5}$$

where $\Psi(A_m) = \frac{2}{N_m(N_m-1)} \sum_{i<j} L(A_m; \mathbf{x}_{mi}, \mathbf{x}_{mj}, y_{mij})$ is the empirical loss w.r.t. A_m in the m'th domain, and $R(A_1, A_2, \ldots, A_M)$ is some regularizer to enforce information transfer across different domains. Following [Jin *et al.*, 2009], we choose $L(A_m; \mathbf{x}_{mi}, \mathbf{x}_{mj}, y_{mij}) = g(y_{mij}[1 - \|\mathbf{x}_{mi} - \mathbf{x}_{mj}\|_{A_m}^2])$ and adopt the hinge loss for g, i.e., $g(z) = \max(0, b - z)$. Here, b is set to be zero, and $\|\mathbf{x}_{mi} - \mathbf{x}_{mj}\|_{A_m}^2 = (\mathbf{x}_{mi} - \mathbf{x}_{mj})^T A_m (\mathbf{x}_{mi} - \mathbf{x}_{mj})$. For notation simplicity, we denote $\mathbf{x}_{mi}$, $\mathbf{x}_{mj}$ and y_{mij} as $\mathbf{x}_{mk}^1$, $\mathbf{x}_{mk}^2$ and y_{mk} respectively, where $k = 1, 2, \ldots, N'_m = \frac{N_m(N_m-1)}{2}$. We also set $\delta_{mk} = \mathbf{x}_{mk}^1 - \mathbf{x}_{mk}^2$ so that $\|\mathbf{x}_{mk}^1 - \mathbf{x}_{mk}^2\|_{A_m}^2 = \delta_{mk}^T A_m \delta_{mk}$, and the loss term becomes $\Psi(A_m) = \frac{1}{N'_m} \sum_{k=1}^{N'_m} g\left(y_{mk}(1 - \delta_{mk}^T A_m \delta_{mk})\right)$.

To enable knowledge transfer across domains, we propose to decompose the positive semi-definite metric A_m as $A_m = U_m U_m^T$, and then using the feature mapping $U_m \in \mathbb{R}^{d_m \times r}$ to project the unlabeled data points of different domains into a common subspace, where the correlation of all domains are maximized. This leads to the following optimization problem:

$$\mathop{\arg\max}_{\{U_m\}_{m=1}^M} \frac{1}{N^U} \sum_{n=1}^{N^U} \mathrm{corr}(\mathbf{z}_{1n}^U, \mathbf{z}_{2n}^U, \ldots, \mathbf{z}_{Mn}^U), \tag{6}$$

where $\mathrm{corr}(\mathbf{z}_{1n}^U, \mathbf{z}_{2n}^U, \ldots, \mathbf{z}_{Mn}^U) = (\mathbf{z}_{1n}^U \odot \mathbf{z}_{2n}^U \odot \ldots \odot \mathbf{z}_{Mn}^U)^T \mathbf{e}$ is the correlation of the projected representations $\{\mathbf{z}_{mn}^U = U_m^T \mathbf{x}_{mn}^U\}_{m=1}^M$ among all domains for the n'th sample. Here, $\odot$ is the element-wise product, and $\mathbf{e} \in \mathbb{R}^r$ is an all ones vector. This correlation is equivalent to $\mathcal{G} \bar{\times}_1 (\mathbf{x}_{1n}^U)^T \ldots \bar{\times}_M (\mathbf{x}_{Mn}^U)^T$ according to [Luo *et al.*, 2015], where $\mathcal{G} = \sum_{q=1}^r (\mathbf{u}_1^q \circ \mathbf{u}_2^q \circ \ldots \circ \mathbf{u}_M^q) = \mathcal{I}_r \times_1 U_1 \times_2 U_2 \ldots \times_M U_M$ is the covariance tensor of the mappings. Here, $\circ$ is the outer product, $\mathcal{I}_r \in \mathbb{R}^{r \times r \times \ldots \times r}$ is an identity tensor (the entries are 1 in the diagonal, and 0 otherwise) of size r, which is the number of common factors shared by all domains. Then the problem (6) becomes

$$\mathop{\arg\max}_{\{U_m\}_{m=1}^M} \frac{1}{N^U} \sum_{n=1}^{N^U} \mathcal{G} \bar{\times}_1 (\mathbf{x}_{1n}^U)^T \ldots \bar{\times}_M (\mathbf{x}_{Mn}^U)^T, \tag{7}$$

Let $\mathcal{C}_n^U = \mathbf{x}_{1n}^U \circ \mathbf{x}_{2n}^U \circ \ldots \circ \mathbf{x}_{Mn}^U$ be the covariance tensor of the original feature representations among all domains for the n'th sample, the above problem can be further reformulated as follows according to [De Lathauwer *et al.*, 2000a],

$$\mathop{\arg\min}_{\{U_m\}_{m=1}^M} \frac{1}{N^U} \sum_{n=1}^{N^U} \|\mathcal{C}_n^U - \mathcal{G}\|_F^2. \tag{8}$$

By regarding the objective of (8) as the regularizer in (5), we obtain the following specific optimization problem for HMTML:

$$\begin{aligned} &\mathop{\arg\min}_{\{U_m\}_{m=1}^M} F(\{U_m\}) \\ &= \sum_{m=1}^M \frac{1}{N'_m} \sum_{k=1}^{N'_m} g\left(y_{mk}(1 - \delta_{mk}^T U_m U_m^T \delta_{mk})\right) \\ &\quad + \frac{\gamma}{N^U} \sum_{n=1}^{N^U} \|\mathcal{C}_n^U - \mathcal{G}\|_F^2 + \sum_{m=1}^M \gamma_m \|U_m\|_1, \\ &\text{s.t. } U_m \succeq 0, m = 1, 2, \ldots, M, \end{aligned} \tag{9}$$

where γ and $\{\gamma_m\}$ are all positive tradeoff parameters. We enforce the feature mapping to be sparse as suggested in [Zhou *et al.*, 2014] and the non-negativity constraints are to preserve non-negative correlation between the original feature representations. Intuitively, minimization of the second term in (9) corresponds to find a latent subspace where the representations of all domains are close to each other. Knowledge is transferred in this subspace and so different domains can help each other in learning the mapping U_m, or equivalently the metric A_m.

2.3 Optimization Algorithm

The problem (9) can be solved by iteratively updating only one variable U_m at a time and fixing all the other U_m, $m' \neq m$. According to [De Lathauwer *et al.*, 2000b], we have

$$\mathcal{G} = \mathcal{I}_r \times_1 U_1 \times_2 U_2 \ldots \times_M U_M = \mathcal{B} \times_m U_m.$$

where $\mathcal{B} = \mathcal{I}_r \times_1 U_1 \ldots \times_{m-1} U_{m-1} \times_{m+1} U_{m+1} \ldots \times_M U_M$. By applying the metricizing property of the tensor-matrix product, we have $G_{(m)} = U_m B_{(m)}$. Besides, it is easy to verify that $\|\mathcal{C}_n^U - \mathcal{G}\|_F^2 = \|C_{n(m)}^U - G_{(m)}\|_F^2$. Therefore, the sub-problem of (9) w.r.t. U_m becomes:

$$\begin{aligned} &\arg\min_{U_m} F(U_m) = \Phi(U_m) + \Omega(U_m), \\ &\quad \text{s.t. } U_m \succeq 0, \end{aligned} \tag{10}$$

where $\Phi(U_m) = \frac{1}{N'_m} \sum_{k=1}^{N'_m} g\left(y_{mk}(1 - \delta_{mk}^T U_m U_m^T \delta_{mk})\right) + \gamma_m \|U_m\|_1$, and $\Omega(U_m) = \frac{\gamma}{N^U} \sum_{n=1}^{N^U} \|C_{n(m)}^U - U_m B_{(m)}\|_F^2$. We propose to solve the problem (10) efficiently by utilizing the projected gradient method (PGM) presented in [Lin, 2007]. However, the terms in $\Phi(U_m)$ are non-differentiable, we thus first smooth it according to [Nesterov, 2005]. For notation clarity, we omit the subscript m in the following derivation. According to [Nesterov, 2005], the smoothed version of the hinge loss $g(U; \delta_k, y_k) = \max\{0, -y_k(1 - \delta_k^T U U^T \delta_k)\}$ can be given by

$$g^\sigma = \max_{\nu \in \mathcal{Q}} \nu_k \left(-y_k(1 - \delta_k^T U U^T \delta_k)\right) - \frac{\sigma}{2} \|\delta_k\|_\infty \nu_k^2, \tag{11}$$

where $\mathcal{Q} = \{\nu : 0 \leq \nu_k \leq 1, \nu \in \mathbb{R}^{N'}\}$ and σ is the smooth parameter, which is set as 0.5 in this paper. By setting the gradient of the objective function in (11) to become zero and then projecting ν_k on $\mathcal{Q}$, we obtain the following solution,

$$\nu_k = \text{median}\{\frac{-y_k(1 - \delta_k^T U U^T \delta_k)}{\sigma \|\delta_k\|_\infty}, 0, 1\}. \tag{12}$$

By substituting the solution (12) back into (11), we have the piece-wise approximation of g, i.e.,

$$g^\sigma = \begin{cases} 0, & \begin{array}{r} y_k(1 - \delta_k^T U U^T \delta_k) \\ > 0 \end{array}; \\ y_k(\delta_k^T U U^T \delta_k - 1) - \frac{\sigma}{2}\|\delta_k\|_\infty, & \begin{array}{r} y_k(1 - \delta_k^T U U^T \delta_k) \\ < -\sigma\|\delta_k\|_\infty \end{array}; \\ \frac{(y_k(1 - \delta_k^T U U^T \delta_k))^2}{2\sigma\|\delta_k\|_\infty}, & \text{otherwise.} \end{cases} \tag{13}$$

To utilize the PGM for optimization, we have to compute the gradient of the smoothed hinge loss to determine the descent direction. We summarize the results in the following theorem.

Theorem 1. *The sum of gradient of the smoothed hinge loss* $g^\sigma(U; \delta_k, y_k)$ *over all samples is*

$$\frac{\partial g^\sigma(U)}{\partial U} = \sum_k (2 y_k \nu_k (\delta_k \delta_k^T) U). \tag{14}$$

Here, ν_k *is related to* U.

It is easy to prove this theorem according to (12) and (13), so we do not present the proof here due to the limited space. Similarly, for the sparse term $\|U\|_1 = \sum_{i=1}^d \sum_{j=1}^r l(u_{ij})$, where $l(u_{ij}) = |u_{ij}|$, we have the following piece-wise approximation of l with the smooth parameter σ:

$$l^\sigma = \begin{cases} -u_{ij} - \frac{\sigma}{2}, & u_{ij} < -\sigma; \\ u_{ij} - \frac{\sigma}{2}, & u_{ij} > \sigma; \\ \frac{u_{ij}^2}{2\sigma}, & \text{otherwise.} \end{cases} \tag{15}$$

The gradient of smoothed $\|U\|_1$ is given by $\partial(\sum_{i=1}^d \sum_{j=1}^r l^\sigma(u_{ij}))/\partial U = O$ with each $o_{ij} = \text{median}\{\frac{u_{ij}}{\sigma}, -1, 1\}$. In addition, it is easy to deduce that the gradient of $\Omega(U)$ w.r.t. U is

$$\frac{\partial \Omega(U)}{\partial U} = \frac{2\gamma}{N^U} \sum_n (U B B^T - C_n^U B^T). \tag{16}$$

Therefore, the gradient of the smoothed $F(U_m)$ is

$$\begin{aligned} \frac{\partial F^\sigma(U_m)}{\partial U_m} &= \frac{1}{N'_m} \sum_k \left(2 y_{mk} \nu_{mk} (\delta_{mk} \delta_{mk}^T) U_m\right) \\ &+ \frac{2\gamma}{N^U} \sum_n \left(U_m B_{(m)} B_{(m)}^T - C_{n(m)}^U B_{(m)}^T\right) + \gamma_m O_m, \end{aligned} \tag{17}$$

Finally, based on the obtained gradient, we apply the improved PGM presented in [Lin, 2007] to minimize the smoothed primal $F^\sigma(U_m)$, i.e.,

$$U_m^{t+1} = P[U_m^t - \mu_t \nabla F^\sigma(U_m^t)], \tag{18}$$

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

$$F^\sigma(U_m^{t+1}) - F^\sigma(U_m^t) \leq \kappa \nabla F^\sigma(U_m^t)^T (U_m^{t+1} - U_m^t), \tag{19}$$

where the parameter κ is chosen to be 0.01 following [Lin, 2007]. The step size can be determined using the Algorithm 4 in [Lin, 2007], and the convergence of the algorithm is guaranteed according to [Lin, 2007]. The stopping criterion we utilized here is $|F^\sigma(U_m^{t+1}) - F^\sigma(U_m^t)|/(|F^\sigma(U_m^{t+1}) - F^\sigma(U_m^0)| < \epsilon)$, where the initialization U_m^0 is the set as the results of the previous iterations in the alternating of all $\{U_m\}_{m=1}^M$.

Finally, the solutions of (9) are obtained by alternatively updating each U_m until the stop criterion $|OBJ_{k+1} - OBJ_k|/|OBJ_k| < \epsilon$ is reached, where OBJ_k is the objective value of (9) in the k'th iteration step. Because the objective value of (10) decreases at each iteration of the alternating procedure, i.e., $F(U_m^{k+1}, \{U_{m'}^k\}_{m' \neq m}) \leq F(\{U_m^k\})$. This indicates that $F(\{U_m^{k+1}\}) \leq F(\{U_m^k\})$. Therefore, the convergence of the proposed HMTML algorithm is guaranteed. Once the solutions $\{U_m^*\}_{m=1}^M$ have been obtained, we can conduct subsequent learning, such as multi-class classification in each domain using the learned metric $A_m^* = U_m^* {U_m^*}^T$.

3 Experiments

In this section, we evaluate the effectiveness of the proposed HMTML on both multi-lingual document categorization and multi-view image annotation. Prior to these evaluations, we present the experimental settings.

3.1 Datasets, Features, and Evaluation Criteria

The dataset used in document categorization is the Reuters multilingual collection (RMLC) [Amini *et al.*, 2009], which contains news articles written in five languages, and from six populous categories. In this dataset, we choose three languages (i.e., English (EN), Italian (IT), and Spanish (SP)) and regard each of them as a domain. The provided TF-IDF features are adopted for document representation. We preprocess these representations by performing PCA to find comparable patterns for meaningful transfer and 20% energy is preserved. This results in 245, 213, and 107 features for documents of the three domains respectively. The number of samples for the three domains are $18,758$, $24,039$, and $12,342$ respectively. In each domain, the sample sets are randomly split into equal size to form the training and test sets, and we randomly choose $\{5, 10, 15\}$ labeled samples for each category in the training set to determine the performance of the compared methods w.r.t. the number of labeled instances.

In image annotation, we employ a challenge natural image dataset NUS-WIDE (NUS) [Chua *et al.*, 2009]. The dataset contains $269,648$ images, and our experiments are conducted on a subset that consists of $16,519$ images belonging to 12 animal concepts: bear, bird, cat, cow, dog, elk, fish, fox, horse, tiger, whale, and zebra. In this dataset, we choose three types of features, namely 500-D bag of visual words (BOVW) based on SIFT [Lowe, 2004] descriptors, 144-D color autocorrelogram (CORR), and 128-D wavelet texture (WT), to represent each image. We preprocess the different features using PCA and the result dimensions are all 100. Each image representation is regarded as a domain. In each domain, we randomly split the image set into a training set of $8,263$ images and a test set of $8,256$ images, and the number of labeled instances for each concept varies in the set $\{4, 6, 8\}$.

In both datasets, the task in each domain is to perform multi-class classification, where the nearest neighbor (1NN) classifier is adopted. 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 all domains are used as unlabeled data. The parameters are determined using leave-one-out cross validation on the labeled set. The classification accuracy is utilized as evaluation criteria. The average performance of all domains is calculated for comparison. In all the following experiments, five random choices of the labeled instances are used, and the mean values are reported.

3.2 Experimental Results and Analysis

The comparison baselines are listed as below:

- **EU:** directly computing the Euclidean distance between samples based on their original feature representations in each domain.
- **RDML [Jin *et al.*, 2009]:** learning the distance metric for each domain separately using the efficient and competitive regularized distance metric learning algorithm presented in [Jin *et al.*, 2009]. This method only utilizes the given limited labeled samples in each domain, and does not make use of any additional information from other domains. The trade-off parameter is chosen from the set $\{10^i | i = -5, -4, \ldots, 4\}$.
- **DAMA [Wang and Mahadevan, 2011]:** constructing mappings U_m to link multiple heterogeneous domains using manifold alignment. The parameter is determined according to the strategy presented in [Wang and Mahadevan, 2011].
- **MTDA [Zhang and Yeung, 2011]:** performing supervised dimension reduction simultaneously for heterogeneous features (domains) using the multi-task extension of linear discriminative analysis. The learned transformation $U_m = W_m P$, which consists of a domain specific part W_m, and a common part P shared by all domains. The intermediate dimensionality parameter is set as 100 for both datasets since the model is not very sensitive to the parameter according to [Zhang and Yeung, 2011].
- **HMTML:** the proposed heterogeneous multi-task metric learning method. The parameters γ_m are set as the same value, and we tune both γ and γ_m over the set $\{10^i | i = -5, -4, \ldots, 4\}$.

In DAMA and MTDA, after learning U_m, we derive the metric for each domain as $A_m = U_m U_m^T$. For DAMA, MTDA, and the proposed HMTML, the number r of the common factors (or dimensionality of the common subspace) used to explain the original data of all domains varies in $\{1, 2, 5, 8, 10, 20, 30, 50, 80, 100\}$.

Multilingual Document Categorization

The classification accuracies in relation to the number r are shown in Figure 2. From these results, we observe that: 1) the performance of all the compared methods improves with an increased number of labeled instances; 2) although the labeled samples in each domain is scarce, learning the distance metric separately using RDML can still improve the performance significantly. This demonstrates the effectiveness of distance metric learning (DML) in this application; 3) all the three heterogeneous transfer learning approaches achieve much better performance than RDML. This indicates that it is useful to leverage information from other domains in DML; Besides, the optimal number r is usually less than 30. This can be interpreted as using only 30 common factors (topics) is enough to distinguish the different categories in the dataset; 4) DAMA is superior to MTDA when the number of labeled samples is small, since MTDA is a discriminative method and highly relies on the label information, while DAMA preserves topology in each domain and this is helpful given insufficient labeled instances. The proposed HMTML is superior to DAMA even limited labeled samples are provided, since we make use of large amounts of unlabeled data to connect different domains; 5) overall, the proposed HMTML outperforms both DAMA and MTDA at most numbers (of common factors). This indicates that the learned factors by our method are more expressive than the other approaches. The main reason is that our method directly examining the high-order statistics of all domains simultaneously, whereas, in DAMA only the pairwise relationships are explored, and in MTDA the different domains must communicate with each other through

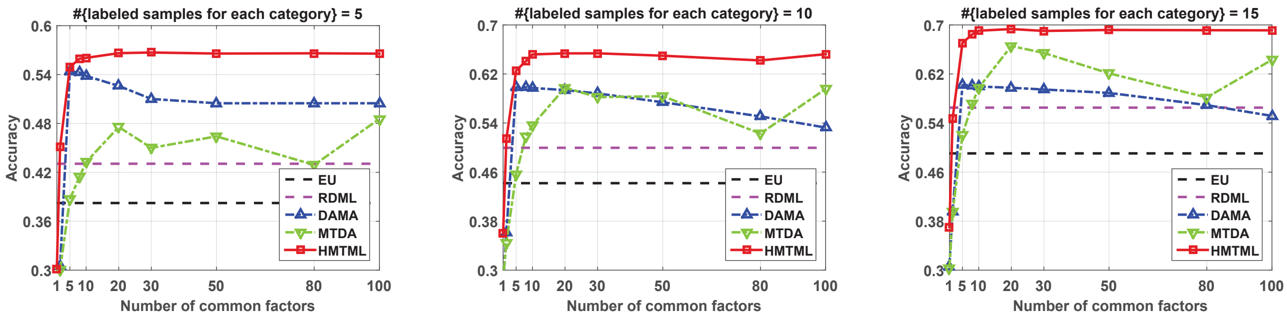

Figure 2: Average accuracy of all domains vs. number of the common factors on the RMLC dataset.

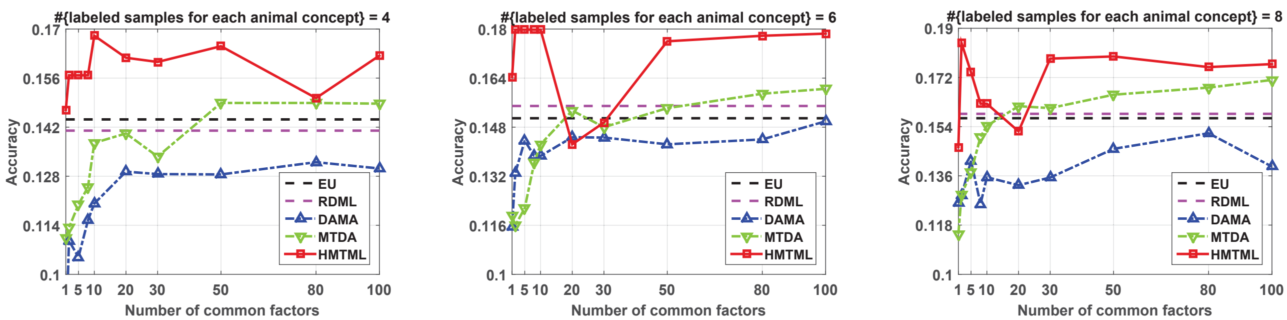

Figure 3: Average accuracy of all domains vs. number of the common factors on the NUS animal subset.

an intermediate structure, where some important information contained in the original features may be lost; 6) in particular, we obtain significant relative improvements of 16.8%, 9.4%, and 4.2% over the competitive MTDA when the number of labeled samples are 5, 10, and 15 respectively.

Multi-view Image Annotation

We show the annotation accuracies of the compared methods in Figure 3. It can be observed from the results that: 1) the accuracy of RDML is lower than directly using the Euclidean distance (EU) when the number of labeled samples is small (e.g., 4). This may be because RDML is a linear metric learning approach, while structure of the data distribution of image features is usually nonlinear; 2) DAMA totally fails in this application, and MTDA only obtain satisfactory accuracies when enough (e.g., 8) labeled instances are provided. The main reason is that in this application, the different domains corresponding to different kinds of features. This setting is much more challenge than the multilingual document classification, where the feature types (TF-IDF) are the same and only the vocabulary varies. The statistical properties of the different kinds visual features utilized here are quite different from each other, so it is very hard to find some common expressive factors across all domains by only exploiting the pair-wise relationships between them. Nevertheless, the proposed HMTML achieves satisfactory performance by simultaneously exploring all domains.

4 Conclusion

This paper presents a method for heterogeneous metric learning. The proposed method can not only effectively make use of the limited side information in each domain, but also discover high order statistics among multiple heterogeneous domains by analyzing their feature covariance tensor calculated using large amounts of unlabeled data. The knowledge shared by the different domains is successfully transferred in a common subspace to help each of them in metric learning by maximizing their high-order covariance in the subspace. We develop an efficient algorithm for optimization with convergence guarantee, and the exploited high-order correlation information was demonstrated empirically to be superior to the pairwise correlations utilized in traditional approaches.

From the experimental validation on two popular applications we mainly conclude that: 1) learning metric for each domain separately may deteriorate the performance if given insufficient side information, and the labeled data deficiency problem can be alleviated by learning metrics for multiple heterogeneous domains simultaneously. This is consistent with the results of multi-task learning literatures; 2) the shared knowledge of different domains exploited by the transfer learning methods can benefit each domain if appropriate common factors are discovered, and the high-order statistics (correlation information) is critical in discovering such factors; In the future, we plan to extend the proposed method to learn nonlinear metrics so that it has the capability to handle complicated domains.

Acknowledgments

This work is supported by Singapore MOE Tier 2 (ARC42/13).

References

[Amini *et al.*, 2009] Massih Amini, Nicolas Usunier, and Cyril Goutte. Learning from multiple partially observed views-an application to multilingual text categorization. In *Advances in Neural Information Processing Systems*, pages 28–36, 2009.

[Ammar *et al.*, 2015] Haitham Bou Ammar, Eric Eaton, José Marcio Luna, and Paul Ruvolo. Autonomous cross-domain knowledge transfer in lifelong policy gradient reinforcement learning. In *International Joint Conference on Artificial Intelligence*, pages 3345–3351, 2015.

[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 *International Conference on Image and Video Retrieval*, 2009.

[De Lathauwer *et al.*, 2000a] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[De Lathauwer *et al.*, 2000b] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. On the best rank-1 and rank-(r 1, r 2,..., rn) approximation of higher-order tensors. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1324–1342, 2000.

[Goetschalckx *et al.*, 2015] Robby Goetschalckx, Alan Fern, and Prasad Tadepalli. Multitask coactive learning. In *International Joint Conference on Artificial Intelligence*, pages 3518–3524, 2015.

[Jin *et al.*, 2009] Rong Jin, Shijun Wang, and Yang Zhou. Regularized distance metric learning: Theory and algorithm. In *Advances in Neural Information Processing Systems*, pages 862–870, 2009.

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

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

[Lowe, 2004] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[Luo *et al.*, 2013] Yong Luo, Dacheng Tao, Bo Geng, Chao Xu, and Stephen J Maybank. Manifold regularized multitask learning for semi-supervised multilabel image classification. *IEEE Transactions on Image Processing*, 22(2):523–536, 2013.

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

[Luo *et al.*, 2015] Yong Luo, Dacheng Tao, Kotagiri Ramamohanarao, Chao Xu, and Yonggang Wen. Tensor canonical correlation analysis for multi-view dimension reduction. *IEEE Transactions on Knowledge and Data Engineering*, 27(11):3111–3124, 2015.

[Luo *et al.*, 2016] Yong Luo, Yonggang Wen, Dacheng Tao, Jie Gui, and Chao Xu. Large margin multi-modal multi-task feature extraction for image classification. *IEEE Transactions on Image Processing*, 25(1):414–427, 2016.

[McFee and Lanckriet, 2010] Brian McFee and Gert R Lanckriet. Metric learning to rank. In *International Conference on Machine Learning*, pages 775–782, 2010.

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

[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 *SIAM International Conference on Data Mining*, pages 528–539, 2012.

[Shi *et al.*, 2010] Xiaoxiao Shi, Qi Liu, Wei Fan, Philip S Yu, and Ruixin Zhu. Transfer learning on heterogenous feature spaces via spectral transformation. In *International Conference on Data Mining*, pages 1049–1054, 2010.

[Wang and Mahadevan, 2011] Chang Wang and Sridhar Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *International Joint Conference on Artificial Intelligence*, pages 1541–1546, 2011.

[Xing *et al.*, 2002] Eric P Xing, Michael I Jordan, Stuart Russell, and Andrew Ng. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, pages 505–512, 2002.

[Xu *et al.*, 2013] Zhixiang Xu, Kilian Q Weinberger, and Olivier Chapelle. Distance metric learning for kernel machines. *arXiv preprint arXiv:1208.3422v2*, 2013.

[Zha *et al.*, 2009] Zheng-Jun Zha, Tao Mei, Meng Wang, Zengfu Wang, and Xian-Sheng Hua. Robust distance metric learning with auxiliary knowledge. In *International Joint Conference on Artificial Intelligence*, pages 1327–1332, 2009.

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

[Zhang and Yeung, 2012] Yu Zhang and Dit-Yan Yeung. Transfer metric learning with semi-supervised extension. *ACM Transactions on Intelligent Systems and Technology*, 3(3):54:1–54:28, 2012.

[Zhou *et al.*, 2014] Joey Tianyi Zhou, Ivor W Tsang, Sinno Jialin Pan, and Mingkui Tan. Heterogeneous domain adaptation for multiple classes. In *International Conference on Artificial Intelligence and Statistics*, pages 1095–1103, 2014.