Exploiting High-Order Information in Heterogeneous Multi-Task Feature Learning

Yong Luo[†], Dacheng Tao[‡], Yonggang Wen[†]
[†]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, dacheng.tao@sydney.edu.au, ygwen@ntu.edu.sg

Abstract

Multi-task feature learning (MTFL) aims to improve the generalization performance of multiple related learning tasks by sharing features between them. It has been successfully applied to many pattern recognition and biometric prediction problems. Most of current MTFL methods assume that different tasks exploit the same feature representation, and thus are not applicable to the scenarios where data are drawn from heterogeneous domains. Existing heterogeneous transfer learning (including multi-task learning) approaches handle multiple heterogeneous domains by usually learning feature transformations across different domains, but they ignore the high-order statistics (correlation information) which can only be discovered by simultaneously exploring all domains. We therefore develop a tensor based heterogeneous MTFL (THMTFL) framework to exploit such high-order information. Specifically, feature transformations of all domains are learned together, and finally used to derive new representations. A connection between all domains is built by using the transformations to project the pre-learned predictive structures of different domains into a common subspace, and minimizing their divergence in the subspace. By exploring the high-order information, the proposed THMTFL can obtain more reliable feature transformations compared with existing heterogeneous transfer learning approaches. Extensive experiments on both text categorization and social image annotation demonstrate superiority of the proposed method.

1 Introduction

Multi-task learning (MTL) [Caruana, 1997; Liu *et al.*, 2017] learns multiple related tasks simultaneously. It aims to choose an appropriate hypothesis space for each learning task by utilizing the shared information across all tasks, and thus improves their generalization performance. A dozen MTL methods [Luo *et al.*, 2013; Ammar *et al.*, 2015; Wang *et al.*, 2016; Luo *et al.*, 2016; Gonçalves *et al.*, 2017] have been proposed in the past decades. For example, a common representation is learned in [Caruana, 1997] for multiple similar tasks by sharing hidden units in neural networks. In [Ammar *et al.*, 2015], some latent factors are assumed to be shared across multiple tasks in lifelong reinforcement learning. The MTL methods can be divided into several groups, and one representative group is multi-task feature learning (MTFL) [Argyriou *et al.*, 2008], which improves the performance of each task by learning some shared features between tasks. It has been widely applied to many real-world applications, such as handwritten letter recognition [Liu *et al.*, 2009], spoken language recognition [Gong *et al.*, 2013], disease diagnosis [Argyriou *et al.*, 2008], and medical state examination [Gong *et al.*, 2013].

An implicit assumption of most existing MTFL algorithms is that the data samples of related domains have the same feature dimensionality or lie in the same feature space, and thus the same feature mapping is learned for all tasks. This assumption may be not valid for many applications. For example, in multilingual document classification, the feature representations of the documents written in different languages vary since the utilized vocabularies are different. In natural image classification and multimedia retrieval, the instances are often represented using different types of features [Xu *et al.*, 2014; 2015] (such as local SIFT [Lowe, 2004] and global wavelet texture) or have different modalities (such as image, audio and text).

Recently, heterogeneous transfer learning (including MTL and domain adaptation) has been proposed to manage heterogeneous representations. These approaches often transform the heterogeneous features into a common subspace, so that the difference between heterogeneous domains is reduced. Although effective in some cases, most of them are limited for only two domains (one source and one target domain). Only a few approaches [Wang and Mahadevan, 2011; Zhang and Yeung, 2011; Jin *et al.*, 2015] could learn transformations for more than two domains. In these approaches, only the statistics (correlation information) between each representation and the shared representation [Zhang and Yeung, 2011; Jin *et al.*, 2015], or pairs of representations [Wang and Mahadevan, 2011] is explored, while high-order statistics (correlation information) [Tao *et al.*, 2007a; 2007b] that can only be obtained by simultaneously examining all domains is ignored.

To deal with an arbitrary number of domains and exploit the high-order information, we develop a tensor based heterogeneous multi-task feature learning (THMTFL) framework. In particular, THMTFL learns the feature mappings for all domains in a single optimization problem. Predictive structures of the different domains are pre-learned to enable knowledge transfer. Following the multi-task learning strategies presented in [Ando and Zhang, 2005; Quattoni *et al.*, 2007], we project the predictive structures of different domains into a common subspace using the feature mappings. In this paper, we assume the application tasks in different domains are the same [Wang and Mahadevan, 2011], such as to classify an article into one of several predefined categories. Consequently, the predictive structures (parameterized by the weight vectors of classifiers) should be close to each other in the subspace. By minimizing the divergence between the transformed predictive structures, we build a connection between the transformations. The label information (included in the predictive structures) of all domains is transferred in the subspace to help learning the transformation (feature mapping) of each domain. Hence the learned feature mappings are more discriminative than the results of learning them separately, especially for those domains that have limited label information.

Our method is superior to existing heterogeneous transfer learning approaches in that: 1) we aim to directly explore the relationship between all domains by analyzing the high-order covariance tensor over their prediction weights. The high-order correlations can thus be encoded in the learned transformations, and hopefully better performance can be achieved; 2) the predictive structures can be learned beforehand, and thus the complexity of learning the feature mapping is reduced. The original data can be invisible in the feature learning. Hence the proposed method can be used in applications where original data are not available due to privacy or security reasons. We perform experiments on two popular applications: text categorization and social image annotation. The results validate the superiority of the proposed THMTFL.

2 Tensor Based Heterogeneous Multi-task Feature Learning

In contrast to the traditional multi-domain heterogeneous transfer learning approaches [Wang and Mahadevan, 2011; Zhang and Yeung, 2011; Jin *et al.*, 2015], which learn linear transformation for each domain by only considering the pairwise correlations, we propose tensor based heterogeneous MTFL (THMTFL) to learn transformations by exploiting the high-order tensor correlation between all domains. Given M heterogeneous domains (such as “English”, “Italian”, and “German” in multilingual document classification), we assume there are limited labeled samples for each of them. For the m'th domain, we construct multiple prediction (such as binary classification) problems and use the labeled data to learn a set of prediction weight vectors $\{\mathbf{w}_m^p\}_{p=1}^P$, where P is the number of prediction problems. Here, each $\mathbf{w}_m^p$ is learned by assuming the prediction base classifier for each domain is linear, i.e., $f^p(\mathbf{x}_m) = (\mathbf{w}_m^p)^T \mathbf{x}_m$. The P prediction problems are generated using the Error Correcting Output Codes (ECOC) scheme [Dietterich and Bakiri, 1995]. Then we use the feature mapping U_m to transform the weight vectors into a common subspace as $\{\mathbf{v}_m^p = U_m^T \mathbf{w}_m^p\}_{p=1}^P$. Because the application tasks in all domains are the same (as mentioned in the Introduction), the transformed weight vectors of different domains $\{\mathbf{v}_1^p, \mathbf{v}_2^p, \ldots, \mathbf{v}_m^p\}$ should be close to each other in the subspace. Finally, by minimizing the tensor based high-order divergence (or equivalently maximizing high-order covariance [Luo *et al.*, 2015]) between all transformed weight vectors, we learn improved U_m^* by utilizing additional information from other domains.

It should be noted that the learned weights may be not reliable given the limited labeled samples. Fortunately, we can construct sufficient base classifiers using the ECOC scheme. Hence robust transformations can be obtained even some learned base classifiers are inaccurate or incorrect. The technical details of the proposed method are given below, and we start by briefing the frequently 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., $B_{(m)} = U A_{(m)}$. Specifically, the series of m-mode product in (2) can be expressed as a series of Kronecker products and is given by

$$B_{(m)} = U_m A_{(m)} \left(U^{(c_{m-1})} \otimes U^{(c_{m-2})} \otimes \ldots \otimes U^{(c_1)} \right)^T, \tag{3}$$

where $\{c_1, c_2, \ldots, c_K\} = \{m+1, m+2, \ldots, M, 1, 2, \ldots, m-1\}$ is a forward cyclic ordering for the indices of the tensor dimensions that map to the column of the matrix. 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:

$$\begin{aligned}&\mathcal{B}(i_1, \ldots, i_{m-1}, i_{m+1}, \ldots, i_M)\\ &= \sum\nolimits_{i_m=1}^{I_m} \mathcal{A}(i_1, i_2, \ldots, i_M) \mathbf{u}(i_m).\end{aligned} \tag{4}$$

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

$$\|\mathcal{A}\|_F^2 = \langle \mathcal{A}, \mathcal{A} \rangle = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_M=1}^{I_M} \mathcal{A}(i_1, i_2, \ldots, i_M)^2. \tag{5}$$

2.2 Problem Formulation

Given M heterogeneous domains, we suppose the labeled training set for the m'th domain is given by $\mathcal{D}_m = \{(\mathbf{x}_{mn}, y_{mn})\}_{n=1}^{N_m}$, where $\mathbf{x}_{mn} \in \mathbb{R}^{d_m}$ and its corresponding class label $y_{mn} \in \{1, 2, \ldots, C\}$. Here, we assume that the different domains are about the same application, and thus share the same label set [Wang and Mahadevan, 2011; Jin *et al.*, 2015]. To enable knowledge transfer across domains in the learning of the feature mappings $\{U_m\}$, we first construct P binary classification problems and learn a set of classifiers $\{\mathbf{w}_m^p\}_{p=1}^P$ for each of the M domains using the labeled training data. This learning process can be carried out offline and thus have no impact on the computational cost of subsequent feature mapping learning. Then we propose to transform $\mathbf{w}_m^p$ using the mapping U_m as $\mathbf{v}_m^p = U_m^T \mathbf{w}_m^p$ and minimize the divergence of the transformed weight vectors, i.e., $\{\mathbf{v}_1^p, \mathbf{v}_2^p, ..., \mathbf{v}_M^p\}$. Here, we start from two domains, i.e., $M = 2$, and then generalize the model for $M > 2$.

When $M = 2$, the formulation is given by

$$\min_{U_1, U_2} \frac{1}{2P} \sum_{p=1}^{P} \|\mathbf{v}_1^p - \mathbf{v}_2^p\|_2^2 + \sum_{m=1}^{2} \gamma_m \|U_m\|_1, \quad \text{s.t. } U_1, U_2 \succeq 0, \tag{6}$$

where $U_m \in \mathbb{R}^{d_m \times r}$ is the mapping for the m'th domain, and r is the number of common factors shared by different domains; $\{\gamma_m\}$ are positive trade-off hyper-parameters, the l_1-norm $\|U_m\|_1 = \sum_{i=1}^d \sum_{j=1}^r |u_{mij}|$ is the sum of the absolute values of U_m's elements, and $\succeq$ indicates that each element of U_m is non-negative. We enforce U_m to be sparse since a feature in one domain is usually represented by only a small subset of features in another domain, and the non-negativity constraints are to preserve non-negative correlation between the feature representations.

To generalize (6) for $M > 2$, we reformulate it as

$$\min_{U_1, U_2} \frac{1}{2P} \sum_{p=1}^{P} \|\mathbf{w}_1^p - G\mathbf{w}_2^p\|_2^2 + \sum_{m=1}^{2} \gamma_m \|U_m\|_1, \quad \text{s.t. } U_1, U_2 \succeq 0. \tag{7}$$

Considering that $G = U_1 E_r U_2^T$ and by using the tensor notation, we have $G = E_r \times_1 U_1 \times_2 U_2$, where E_r is an identity matrix of size r. Thus, the formulation (7) for $M > 2$ is:

$$\min_{\{U_m\}_{m=1}^M} \frac{1}{2P} \sum_{p=1}^{P} \|\mathbf{w}_1^p - \mathcal{G} \bar{\times}_2 (\mathbf{w}_2^p)^T \ldots \bar{\times}_M (\mathbf{w}_M^p)^T\|_2^2 + \sum_{m=1}^{M} \gamma_m \|U_m\|_1, \tag{8}$$

where each $U_m \succeq 0$ and $\mathcal{G} = \mathcal{E}_r \times_1 U_1 \times_2 U_2 \ldots \times_M U_M$ is a transformation tensor, $\mathcal{E}_r \in \mathbb{R}^{r \times r \times \ldots \times r}$ is an identity tensor (the entries are 1 in the diagonal, and 0 otherwise). The problem (8) is not convenient to optimize, so we reformulate it by utilizing the following theorem.

Theorem 1 *The following equality holds for* $\|\mathbf{w}_m^p\|_2^2 = 1, p = 1, \ldots, P; m = 1, \ldots, M$*:*

$$\begin{aligned} &\|\mathbf{w}_1^p - \mathcal{G} \bar{\times}_2 (\mathbf{w}_2^p)^T \ldots \bar{\times}_M (\mathbf{w}_M^p)^T\|_2^2 \\ &= \|\mathbf{w}_1^p \circ \mathbf{w}_2^p \ldots \circ \mathbf{w}_M^p - \mathcal{G}\|_F^2, \end{aligned} \tag{9}$$

where $\circ$ *is the outer product.*

Due to limited page length, we omit the proof. By substituting (9) into (8) and replacing $\mathcal{G}$ with $\mathcal{E} \times_1 U_1 \times_2 U_2 \ldots \times_M U_M$, we obtain the following reformulation of (8):

$$\begin{aligned} \min_{\{U_m\}_{m=1}^M} \; & F(\{U_m\}) \\ & = \frac{1}{2P} \sum_{p=1}^{P} \|\mathcal{W}^p - \mathcal{E}_r \times_1 U_1 \times_2 U_2 \ldots \times_M U_M\|_F^2 \\ & + \sum_{m=1}^{M} \gamma_m \|U_m\|_1, \\ & \text{s.t. } U_m \succeq 0, m = 1, 2, \ldots, M, \end{aligned} \tag{10}$$

where $\mathcal{W}^p = \mathbf{w}_1^p \circ \mathbf{w}_2^p \ldots \circ \mathbf{w}_M^p$ is the prediction weights covariance tensor among all domains. Intuitively, minimization of the first term in (10) corresponds to find a subspace where the weight vectors of all domains are close to each other. Knowledge (label information) is transferred in this subspace and thus different domains can help each other in learning the mapping U_m.

2.3 Optimization Algorithm

The problem (10) 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.*, 2000], we have

$$\mathcal{G} = \mathcal{E} \times_1 U_1 \times_2 U_2 \ldots \times_M U_M = \mathcal{B} \times_m U_m,$$

where $\mathcal{B} = \mathcal{E} \times_1 U_1 \ldots \times_{m-1} U_{m-1} \times_{m+1} U_{m+1} \ldots \times_M U_M$ and 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 $\|\mathcal{W}^p - \mathcal{G}\|_F^2 = \|W_{(m)}^p - G_{(m)}\|_F^2$. Therefore, the sub-problem of (10) w.r.t. U_m becomes:

$$\min_{U_m} F(U_m) = \Phi(U_m) + \Omega(U_m), \; \text{ s.t. } U_m \succeq 0, \tag{11}$$

where $\Phi(U_m) = \frac{1}{2P} \sum_{p=1}^P \|W_{(m)}^p - U_m B_{(m)}\|_F^2$, and $\Omega(U_m) = \gamma_m \|U_m\|_1$. We propose to solve the problem (11) efficiently by utilizing the projected gradient method (PGM) presented in [Lin, 2007]. However, the term in $\Omega(U_m)$ is non-differentiable, we thus first smooth it according to [Nesterov, 2005]. For notational clarity, we omit the subscript m in the following derivation. According to [Nesterov, 2005], the smoothed version of the l_1-norm $l(u_{ij}) = |u_{ij}|$ can be

$$l^\sigma(u_{ij}) = \max_{Q \in \mathcal{Q}} \langle u_{ij}, q_{ij} \rangle - \frac{\sigma}{2} q_{ij}^2, \tag{12}$$

where $\mathcal{Q} = \{Q : -1 \leq q_{ij} \leq 1, Q \in \mathbb{R}^{d \times r}\}$ and σ is the smooth hyper-parameter, which is empirically set it as 0.5 in our implementation according to the comprehensive study of the smoothed l_1-norm in [Zhou *et al.*, 2010]. By setting the objective function of (12) to become zero and then projecting q_{ij} on $\mathcal{Q}$, we obtain the following solution:

$$q_{ij} = \text{median}\left\{\frac{u_{ij}}{\sigma}, -1, 1\right\}. \tag{13}$$

By substituting the solution (13) back into (12), we have the piece-wise approximation of l, i.e.,

$$l^{\sigma} = \begin{cases} -u_{ij} - \frac{\sigma}{2}, & u_{ij} < -\sigma; \\ u_{ij} - \frac{\sigma}{2}, & u_{ij} > \sigma; \\ \frac{u_{ij}^2}{2\sigma}, & \text{otherwise.} \end{cases} \tag{14}$$

To utilize the PGM for optimization, we have to compute the gradient of the smoothed l_1-norm to determine the descent direction. It is easy to derive that the gradient $\frac{\partial l^{\sigma}(U)}{\partial U} = Q$, where Q is the matrix defined in (12). Therefore, the gradient of the smoothed $F(U_m)$ is

$$\begin{aligned} \frac{\partial F^{\sigma}(U_m)}{\partial U_m} = &\frac{1}{P}\sum_{p}\left(U_m B_{(m)} B_{(m)}^T - W_{(m)}^p B_{(m)}^T\right) \\ &+ \gamma_m Q_m. \end{aligned} \tag{15}$$

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.,

$$U_m^{t+1} = \pi[U_m^t - \mu_t \nabla F^{\sigma}(U_m^t)], \tag{16}$$

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

$$F^{\sigma}(U_m^{t+1}) - F^{\sigma}(U_m^t) \leq \kappa \nabla F^{\sigma}(U_m^t)^T (U_m^{t+1} - U_m^t), \tag{17}$$

where the hyper-parameter κ 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 $|F^{\sigma}(U_m^{t+1}) - F^{\sigma}(U_m^t)|/(|F^{\sigma}(U_m^{t+1}) - F^{\sigma}(U_m^0)| < \epsilon)$, 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 (10) are obtained by alternatively updating each U_m using Algorithm 1 until the stop criterion $|OBJ_{k+1} - OBJ_k|/|OBJ_k| < \epsilon$ is reached, where OBJ_k is the objective value of (10) in the k'th iteration step. Because the objective value of (11) 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 $F(\{U_m^{k+1}\}) \leq F(\{U_m^k\})$. Consequently, the convergence of the proposed THMTFL 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 transformed features $X_m^* = {U_m^*}^T X_m$.

Algorithm 1 The improved projected gradient method for solving U_m.

```
Input: $B^{(m)}$ and $W_{(m)}^p, p = 1, \ldots, P$.
Algorithm hyper-parameters: $\gamma_m$.
Output: $U_m$.
Set $\mu_0 = 1$, and $\beta = 0.1$. Initialize $U_m^0$ as the results of
  the previous iterations in the alternating optimization of
  all $\{U_m\}_{m=1}^M$.
For $t = 1, 2, \ldots$
    (a) Assign $\mu_t \leftarrow \mu_{t-1}$;
    (b) If $\mu_t$ satisfies (17), repeatedly increase it by
      $\mu_t \leftarrow \mu_t/\beta$, until either $\mu_t$ does not satisfy (17) or
      $U_m(\mu_t/\beta) = U_m(\mu_t)$.
      Else repeatedly decrease $\mu_t \leftarrow \mu_t * \beta$ until $\mu_t$ satisfy
      (17);
    (c) Update $U_m^{t+1} = \pi[U_m^t - \mu_t \nabla F^{\sigma}(U_m^t)]$.
Until convergence
```

In Algorithm 1, suppose the number of checks that is needed to find the step size is T_1, and the number of iterations for reaching the stop criterion is T_2, then the time complexity of the proposed THMTFL is $(\Gamma M[r \prod_{m=1}^M d_m + T_2 T_1 r^2 \bar{d}_m])$, where $\bar{d}_m$ is the average feature dimension of all domains, and Γ is the number of iterations for alternately updating all $\{U_m\}_{m=1}^M$. The complexity is linear w.r.t. M and $\prod_{m=1}^M d_m$, and quadratic in the numbers r. Besides, it is common that $\Gamma < 10$, $T_2 < 20$, and $T_1 < 50$, so the complexity is not very high.

3 Experiments

In this section, we evaluate the effectiveness of the proposed THMTFL on both document categorization and image annotation. Prior to these evaluations, we present the used datasets, evaluation criteria, as well as our 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, Italian, and Spanish) and regard each of them as a domain. The provided TF-IDF features are adopted for document representation. We preprocess these representations by performing principal component analysis (PCA) and this results in 245, 213, and 107 features for documents of the three domains respectively. The preprocessing is mainly to: 1) find comparable and high-level patterns for transfer; 2) pervent overfitting; and 3) reduce computational demands in the experiments. 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. In the training set, we randomly select 10 labeled samples for each category.

In image annotation, we employ a challenging 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 the 500-D local bag of SIFT [Lowe, 2004], 128-D global wavelet texture, and 1000-D tag to represent each image. We preprocess the different features using kernel PCA (KPCA) 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 is 6.

In both datasets, the task in each domain is to perform multi-class classification [Liu and Tao, 2016], where the SVM classifiers are adopted. In all the following experiments, both the classification accuracy and macroF1 [Sokolova and Lapalme, 2009] score are utilized as evaluation criteria. The average performance of all domains is calculated for comparison. Ten random choices of the labeled instances are used, and the mean values with standard deviations are reported.

3.2 Experimental Results and Analysis

The compared methods are listed as below:

- **Original:** directly using the original normalized feature representations in each domain.
- **LDA [Wang *et al.*, 2007]:** learning the feature mapping for each domain separately using the linear (Fisher) discriminant analysis algorithm [Wang *et al.*, 2007]. The algorithm only utilizes the given limited labeled samples in each domain, and does not make use of any additional information from other domains.
- **DAMA [Wang and Mahadevan, 2011]:** constructing mappings $\{U_m\}$ to link multiple heterogeneous domains using manifold alignment. The hyper-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 discriminant analysis. The intermediate dimensionality set as 100 since the model is not very sensitive to the hyper-parameter according to [Zhang and Yeung, 2011].
- **MTNMF [Jin *et al.*, 2015]**: a recently proposed heterogeneous multi-task feature learning algorithm. It learns feature transformations for multiple heterogeneous domains jointly by performing matrix factorizations on their feature-label matrices and sharing the class representations. The trade-off hyper-parameter is optimized over the set $\{10^i|i = -5, -4, \ldots, 4\}$.
- **THMTFL:** the proposed tensor based heterogeneous multi-task feature learning method. Linear SVMs are adopted to learn the weight vectors of base classifiers for knowledge transfer. The hyper-parameters $\{\gamma_m\}$ are set as the same value, and we tune γ_m over the set $\{10^i|i = -5, -4, \ldots, 4\}$. The hyper-parameter P is empirically set as $10\lceil 1.5 \log C \rceil$.

If unspecified, the hyper-parameters are determined using leave-one-out cross validation on the labeled set. For DAMA, MTDA, MTNMF, and the proposed THMTFL, we do not tune the hyper-parameter r, which is the number of common factors (or dimensionality of the common subspace). The performance comparisons are performed on a set of varied $r = \{1, 2, 5, 8, 10, 20, 30, 50, 80, 100\}$. This also applies to the result dimension of LDA.

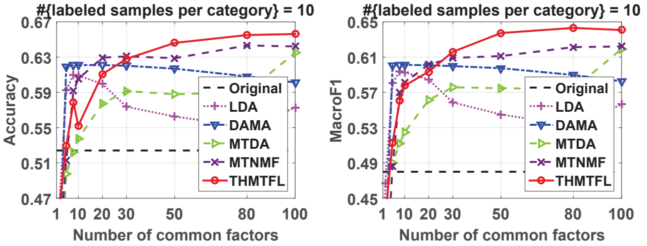

Figure 1: Average accuracy and macroF1 of all domains vs. number of the common factors on the RMLC dataset.

Table 1: Average accuracy and macroF1 score of all domains of the compared methods at their best numbers (of common factors) on the RMLC dataset. (10 labeled instances for each category.)

	Accuracy	MacroF1
Original	0.524±0.021	0.480±0.019
LDA	0.610±0.023	0.594±0.019
DAMA	0.621±0.011	0.601±0.013
MTDA	0.635±0.027	0.619±0.026
MTNMF	0.643±0.021	0.622±0.020
THMTFL	**0.656±0.013**	**0.643±0.008**

Document Categorization

The classification accuracies and macroF1 scores in relation to the number r are shown in Figure 1. The performance of different methods at their best numbers (of common factors) are summarized in Table 1. From these results, we observe that: 1) although the labeled samples in each domain is scarce, learning the feature mapping separately using LDA can still improve the performance significantly. This demonstrates the effectiveness of the supervised mapping learning method in this application; 2) all the heterogeneous transfer learning approaches (DAMA, MTDA, MTNMF, and THMTFL) achieve much better performance than the single domain LDA algorithm. This indicates that it is useful to leverage information from other domains; 3) overall, the proposed THMTFL outperforms all DAMA, MTDA and MTNMF 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. However, in DAMA only the pairwise relationships are explored, and in MTDA and MTNMF the different domains must communicate with each other through an intermediate structure, where some important information contained in the original features may be lost; 4) the performance under the accuracy and macroF1 criteria are consistent. In particular, we obtain a significant relative improvement of 3.4% over the competitive MTNMF in terms of macroF1 score.

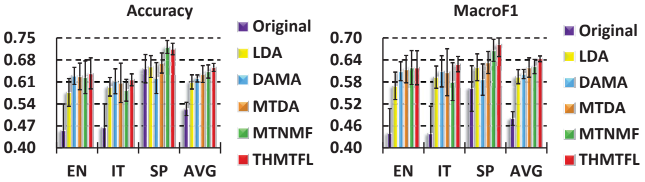

Figure 2: Individual accuracy and macroF1 score of each domain of the compared methods at their best numbers (of common factors) on the RMLC dataset. (10 labeled instances for each category; EN: English, IT: Italian, SP: Spanish, AVG: average.)

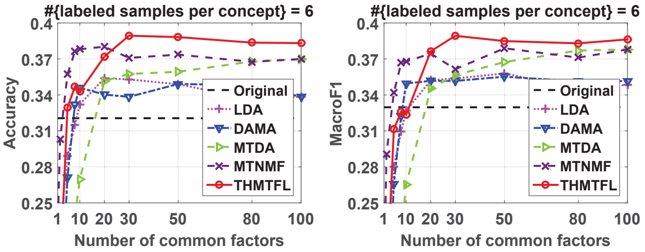

Figure 3: Average accuracy and macroF1 of all domains vs. number of the common factors on the NUS animal subset.

Table 2: Average accuracy and macroF1 score of all domains of the compared methods at their best numbers (of common factors) on the NUS animal subset. (6 labeled instances for each concept.)

	Accuracy	MacroF1
Original	0.321±0.035	0.330±0.029
LDA	0.354±0.015	0.358±0.010
DAMA	0.349±0.017	0.355±0.019
MTDA	0.370±0.015	0.378±0.010
MTNMF	0.380±0.014	0.379±0.012
THMTFL	**0.389±0.006**	**0.389±0.010**

We also show the performance for the individual domains of different methods at their best numbers in Figure 2. It can be seen from the results that: 1) the heterogeneous transfer learning methods has larger improvements than LDA in the domains that the discriminative ability of the original representations is not very good, such as EN (English) and IT (Italian). This demonstrates that the knowledge is successfully transferred between different domains; 2) in the good performing SP (Spanish) domain, MTDA is comparable to LDA and DAMA is even worse than LDA. It seems that the discriminative domain obtains little benefits from the other relatively non-discriminative domains in DAMA and MTDA. Although MTNMF performs well in the SP domain, it fails in the IT domain. The proposed THMTFL achieves satisfactory improvements in all domains. This demonstrates that the high-order correlation information between all domains is well discovered, and that exploring this kind of information is much better than only exploring the correlation information between pairs of domains (as in DAMA) or through an intermediate shared structure (as in MTDA and MTNMF).

Image Annotation

We show the annotation accuracies and MacroF1 scores of the compared methods in Figure 3, and summarize the results at their best numbers (of common factors) in Table 2. It can be observed from the results that: 1) the improvements of LDA compared with the original feature baseline are not that large as in the RMLC dataset. This may be because the data of different concepts in this dataset are less separable than the ones in RMLC. Hence the effectiveness of the supervised feature mapping learning method decreases given the limited label information; 2) DAMA is only comparable to LDA, and the improvements of MTDA compared with LDA are marginal. The main reason is that in this application, the different domains corresponding to different kinds of features. This setting is much more challenging 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. Therefore, it is very hard to find some common expressive factors across all domains by only exploiting the pairwise relationships between them. Nevertheless, the proposed THMTFL achieves satisfactory performance by simultaneously exploring all domains, and is superior to the competitive MTNMF at most numbers (of common factors). This further verifies the superiority of the proposed method. The tendency of the macroF1 score curves are similar to that of the accuracy.

4 Conclusion

This paper presents a method for heterogeneous multi-task learning. The proposed method can discover high-order statistics among multiple heterogeneous domains by analyzing the prediction weight covariance tensor of them. The knowledge shared by the different domains is successfully transferred in a common subspace to help each of them in the feature learning by minimizing their high-order divergence in the subspace. We develop an efficient algorithm for optimization, and the exploited high-order correlation information was demonstrated empirically to be superior to the pairwise correlations utilized in the traditional approaches.

From the experimental validation on two popular applications we mainly conclude that: 1) the labeled data deficiency problem can be alleviated by learning 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 mappings so that it has the capability to handle complicated domains.

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

[Amini *et al.*, 2009] Massih Amini, Nicolas Usunier, and Cyril Goutte. Learning from multiple partially observed views-an application to multilingual text categorization. In *NIPS*, 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 *IJCAI*, pages 3345–3351, 2015.

[Ando and Zhang, 2005] Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *JMLR*, 6:1817–1853, 2005.

[Argyriou *et al.*, 2008] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

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

[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.

[De Lathauwer *et al.*, 2000] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. On the best rank-1 and rank-(r1, r2, ..., rn) approximation of higher-order tensors. *SIAM JMAA*, 21(4):1324–1342, 2000.

[Dietterich and Bakiri, 1995] Thomas G. Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *JAIR*, 2:263–286, 1995.

[Gonçalves *et al.*, 2017] André R Gonçalves, Arindam Banerjee, and Fernando J Von Zuben. Spatial projection of multiple climate variables using hierarchical multitask learning. In *AAAI*, 2017.

[Gong *et al.*, 2013] Pinghua Gong, Jieping Ye, and Changshui Zhang. Multi-stage multi-task feature learning. *JMLR*, 14(1):2979–3010, 2013.

[Jin *et al.*, 2015] Xin Jin, Fuzhen Zhuang, Sinno Jialin Pan, Changying Du, Ping Luo, and Qing He. Heterogeneous multi-task semantic feature learning for classification. In *CIKM*, pages 1847–1850, 2015.

[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.*, 2009] Jun Liu, Shuiwang Ji, and Jieping Ye. Multi-task feature learning via efficient l 2, 1-norm minimization. In *UAI*, pages 339–348, 2009.

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

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

[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 TKDE*, 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 TIP*, 25(1):414–427, 2016.

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

[Quattoni *et al.*, 2007] Ariadna Quattoni, Michael Collins, and Trevor Darrell. Learning visual representations using images with captions. In *IEEE CVPR*, pages 1–8, 2007.

[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.

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

[Wang *et al.*, 2007] Huan Wang, Shuicheng Yan, Dong Xu, Xiaoou Tang, and Thomas Huang. Trace ratio vs. ratio trace for dimensionality reduction. In *IEEE CVPR*, pages 1–8, 2007.

[Wang *et al.*, 2016] Xuezhi Wang, Junier B Oliva, Jeff Schneider, and Barnabás Póczos. Nonparametric risk and stability analysis for multi-task learning problems. In *IJCAI*, pages 2146–2152, 2016.

[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.

[Zhou *et al.*, 2010] Tianyi Zhou, Dacheng Tao, and Xindong Wu. Nesvm: A fast gradient method for support vector machines. In *IEEE ICDM*, pages 679–688, 2010.