Multi-Relevance Transfer Learning

Abstract

Transfer learning aims to faciliate learning tasks in a label-scarce target domain by leveraging knowledge from a related source domain with plenty of labeled data. Often times we may have multiple domains with little or no labeled data as targets waiting to be solved. Most existing efforts tackle target domains separately by modeling the 'source-target' pairs without exploring the relatedness between them, which would cause loss of crucial information, thus failing to achieve optimal capability of knowledge transfer. In this paper, we propose a novel and effective approach called Multi-Relevance Transfer Learning (MRTL) for this purpose, which can simultaneously transfer different knowledge from the source and exploits the shared common latent factors between target domains. Specifically, we formulate the problem as an optimization task based on a collective nonnegative matrix tri-factorization framework. The proposed approach achieves both source-target transfer and target-target leveraging by sharing multiple decomposed latent subspaces. Further, an alternative minimization learning algorithm is developed with convergence guarantee. Empirical study validates the performance and effectiveness of MRTL compared to the state-of-the-art methods.

Introduction

Transfer learning, which intends to utilize knowledge from source domains to help the learning in a target domain, has been established as one of the most important machine learning paradigms (Pan and Yang 2010). In practice, a common scenario is that test data are often sampled from different distributions. One example is the EEG-based Brain Computer Interfaces (BCI) applications. If people want to classify the EEG data collected from several sessions (e.g. more than one hour) while only one session of them are labeled, then the unlabeled sessions can be seen as the target domains and the labeled one is source domain. In this case, the distribution divergences between the source domain and different target domains may vary widely. Another characteristic is that the target domains may share some common latent structure which can help enhance the knowledge transfer. Hence, a significant requirement for sufficient transfer learning in this scenario is to simultaneously exploit the relatedness between target domains and borrow different knowledge from the source domain to each target domain.

However, most existing domain adaptation methods are designed for transferring knowledge from one or multiple source domains to a single target domain. We refer to such approach as single-relevance transfer learning. That is, the information path only between source and target. These methods do not consider the underlying relatedness between target domains. Incurred by the multi-domain property, learning one target domain can help to learn another. It will lead to mutual reinforcement when learning the target domains together. Without exploiting the relatedness between targets, existing methods may only seperately transfer the common knowledge in each 'source-target' pairs, which may result in partial transfer and is difficult to achieve optimal capability of knowledge transfer. To exploit the relatedness between domains, multi-task learning (Evgeniou and Pontil 2007; Dredze *et al.* 2010) is a good choice which tackles these related tasks together by extracting and utilizing appropriate shared information across domains. However, multi-task learning techniques are suitable for the cases that training and test data in each domain are sampled from the same distribution and each domain has reasonably large amounts of labeled data. Therefore, these methods would fail to transfer different knowledge to each target domain from the source, which are not ideal for such applications.

In this paper, we propose a novel approach, Multi-Relevance Transfer Learning (MRTL), which simultaneously transfers different knowledge from the source to each target domain, and exploits the relatedness between targets to achieve knowledge reinforcement. The main idea of MRTL is illustrated in Figure 1. Different from traditional single-relevance methods, MRTL enhances transfer capability by targets exploration. More specifically, MRTL is formulated as an optimization problem of collective nonnegative matrix tri-factorization (NMTF). It decouples the source domain feature into multiple shared latent subspaces as bridges for source-target knowledge transfer and subspaces of remaining feature clusters in each domain. Moreover, the target domains share a cluster association subspace to enable mutual reinforcement. We develop an alternating learning algorithm to optimize the objective. We give theoretical analysis of the proposed algorithm for convergence, and empirically show the effectiveness of the proposed method. Overall, our

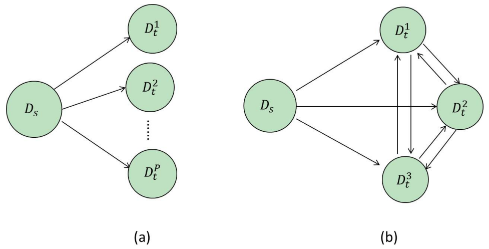


Figure 1: (a) shows traditional single-relevance transfer learning. Knowledge is transferred from source to one target domain each time. (b) is Multi-Relevance Transfer Learning (MRTL). Knowledge simultaneously comes from multiple domains: from source domain and other target domains.

main contributions of this paper include: (1) In addressing multi-relevance transfer learning problem, we propose a M-RTL framework to achieve both source-target transfer and target-target transfer by exploiting shared feature subspaces; (2) We develop an alternating algorithm for optimization; (3) We analyze the theoretical convergence guarantee of the proposed algorithm, and also examine their empirical performances extensively.

Related Work and Preliminaries

In this section, we first discuss several prior researches that mostly related to our work. Then we introduce the NMTF framework as preliminary.

Transfer Learning solves the training data and test data obtained from different resources with different distributions. Most existing efforts assume that there is shared knowledge structure acting as a bridge between the source domain and target domain to enable knowledge transfer. Existing approaches can be grouped into two categories: instance-based transfer learning and feature-based transfer learning (Pan and Yang 2010). Instance-based transfer learning use reweighting strategy to adapting the weights of source domain data (Gretton *et al.* 2009; Jiang and Zhai 2007). The second are feature representation methods which aim to learn a shared feature space to embedding the cross-domain feature information (Blitzer *et al.* 2006; 2011; Pan *et al.* 2009; Gupta *et al.* 2010; Tan *et al.* 2015). Existing feature representation methods focus on transfer single-relevance latent structure from source to target. For example, Dual Transfer Learning (DTL) (Long *et al.* 2012) aims to simultaneously learning the marginal and conditional distributions across domains, Triplex Transfer Learning (TriTL) (Zhuang *et al.* 2014) which make source and target domain share one set of latent subspaces to transfer information. Although their formulations can also be extended to multiple target domains by sharing the same latent structures, it would fail to transfer different knowledge to each target from the source. The key difference between MRTL and these previous methods is that MRTL simultaneously learns different latent subspaces as bridges for knowledge transfer in each 'source-target' pair, which is a crucial step to enhance the transfer diversity and capability.

More recently, multi-source transfer learning (Zhang *et al.* 2015; Duan *et al.* 2012) have been developed to combine knowledge from multiple sources. For instance, the work in (Chattopadhyay *et al.* 2012) present a two-stage domain adaptation method which combines weighted data from multiple sources. The study in (Duan *et al.* 2012) propose a framework which can learn a robust decision function for label prediction for knowledge transfer.

Different from previous transfer learning approaches, multi-relevance transfer learning does not assume that auxiliary knowledge should only come from the source domain. That means, multi-relevance transfer learning can be more general and useful when the existing labeled auxiliary domain is not adequate enough to improve the target domains.

Multi-task Learning approaches simultaneously learn several tasks together to mutual reinforcement the classification results of each task (Obozinski *et al.* 2006; Evgeniou and Pontil 2007; Zhang *et al.* 2012; Dredze *et al.* 2010; Cheng *et al.* 2013). It assumes that different tasks may share some common pattern, such as data clusters or subspaces. In practice, classifiers for different tasks can be designed to share some global parameters (Evgeniou and Pontil 2004) or even a global classifier (Chapelle *et al.* 2010). However, these methods require reasonably large amounts of labeled data in each domain to learn the relationship. In contrast, multi-relevance transfer learning works even when all the learning tasks in each target domain have no available ground truth. It only assumes that the source domain should have sufficient label information to transfer.

Non-negative Matrix Tri-Factorization (NMTF) is popular and effective for data clustering and classification (Ding *et al.* 2006). It can decompose the feature-instance matrix into three submatrices. In general, given feature-instance matrix $X \in \mathbb{R}^{M \times N}$, M is dimensionality and N is instance number. One can obtain the factorized submatrices by solving the optimization problem given by:

$$\min_{U,\Theta,V^{\mathrm{T}}} \mathcal{L} = \|X - U\Theta V^{\mathrm{T}}\|^2, \tag{1}$$

where $\|\cdot\|$ is Frobenius norm of matrix. The matrix $U \in \mathbb{R}^{M \times k}$ indicates *feature cluster subspace* and k is the number of clusters in totoal. U_{ij} is the probability that the i-th feature belongs to the j-th feature cluster. The matrix $V \in \mathbb{R}^{N \times c}$ is the instance *cluster assignment* matrix and c is cluster number. Let $V_{i,\tau} = \max_{1 \leq j \leq c} V_{i,j}$, it means that the i-th instance belongs to the τ-th cluster. For classification, each instance cluster can be regarded as a label class. The matrix $\Theta \in \mathbb{R}^{k \times c}$ is the *cluster association* subspace. Θ_{ij} indicates the probability that the i-th feature cluster is associated with the j-th instance cluster.

Multi-Relevance Knowledge Transfer

We focus on transductive transfer learning where the source domain has abundant labeled examples while the target domains have unlabeled data. We consider one source domain $\mathcal{D}_s$ and multiple target domains $\mathcal{D}_t^p$, $p = 1, 2, ..., P$. $\mathcal{D}_s$ and $\mathcal{D}_t^p$ share the same feature dimensionality and label space. Here we consider M features and c classes. Let $X_s = \left[x_1^s, ..., x_{n_s}^s\right]^{\mathrm{T}} \in \mathbb{R}^{M \times n_s}$ represents the feature-instance matrix of source domain $\mathcal{D}_s$, while $X_t^p = [x_1^p, ..., x_{n_t^p}^p]^{\mathrm{T}} \in \mathbb{R}^{M \times n_t^p}$ denotes the feature-instance matrix of the p-th target domain $\mathcal{D}_t^p$. Labels in the source domain $\mathcal{D}_s$ are given as $Y_s \in \mathbb{R}^{n_s \times c}$, where $y^s_{ij} = 1$ if x_i belongs to class j, and $y^s_{ij} = 0$ otherwise. Given $\{X_s, Y_s\}$ and $\{X^p_t\}^P_{p=1}$, we aim to find a function f to predict the correct label y^p_i for any unlabled instance $x^p_i, i \in [1, n^p_t]$ in the p-th target domain, i.e., $y^p_i = f(x^p_i)$. The goal of *multi-relevance transfer learning* is to alleviate the difficulty of distribution divergences between source-target domains and target-target domains by making them drawn closer in the uncovered latent subspaces so that the classifier f can be trained as accurate as possible.

Model Formulation

To achieve the goal of multi-relevance transfer learning, we propose an algorithm which enable knowledge can be shared between source-target domains and target-target domains for sufficient transfer learning.

Source-Target Knowledge Transfer We first discover the latent factors shared across each 'source-target' pairs. It can be formulated as a collective way of nonnegative matrix trifactorization (NMTF). Given source domain $\mathcal{D}_s$ and the p-th target domain $\mathcal{D}^p_t$, one can decompose their feature-instance matrices X_s and X^p_t simultaneously, allowing the decomposed matrices share the cross-domain latent subspaces. Motivated by (Gupta *et al.* 2010), cross-domain feature clusters can be partitioned into a common part and a domain-specific part. The common feature cluster subspace can be shared across domains, and the domain-specific ones are the remaining feature clusters in each domain. Since we have multiple source-target pairs, we decompose each $\{X_s, X^p_t\}$ as follows:

$$\begin{aligned}\mathcal{L}_1 =& \|X^p_t - [U^p, U^p_t] \left[\begin{array}{c} \Theta^p_\mu \\ \Theta^p_t \end{array} \right] (V^p_t)^{\mathrm{T}} \|^2 \\ &+ \|X_s - [U^p, U^p_s] \left[\begin{array}{c} \Theta^p_\mu \\ \Theta^p_s \end{array} \right] Y^{\mathrm{T}}_s \|^2, \end{aligned} \quad (2)$$

where $U^p \in \mathbb{R}^{M \times k_1}$ is the subspace of *common* feature clusters shared across domains, Θ^p_μ is its corresponding subspace of common cluster association , $U^p_s \in \mathbb{R}^{M \times (k_2 - k_1)}$ and $U^p_t \in \mathbb{R}^{M \times (k_2 - k_1)}$ are the subspaces of *remaining* feature clusters in $\mathcal{D}_s$ and $\mathcal{D}^p_t$ respectively, Θ^p_t and Θ^p_s are the subspaces of remaining cluster association. In this way, the source domain data matrix can be decoupled for each source-target pairs by sharing different subspaces U^p and Θ^p_μ across domain, thus tranferring different knowledge.

Multi-Relevance Transfer Learning Algorithm As shown in Figure 1, in MRTL, knowledge also needs to be transferred between target domains. Since no label information in targets, we need exploit their feature relatedness. That is, we need uncover the shared common latent factors between them as bridges for mutual reinforcement. Therefore, to capture the latent factors between targets, we formulate the factorization of feature-instance matrices of target domains by sharing subspaces as follows:

$$\mathcal{L}_2 = \sum^P_{p=1} \|X^p_t - [U^p, U^p_t] \left[\begin{array}{c} \Theta_\mu \\ \Theta_\sigma \end{array} \right] (V^p_t)^{\mathrm{T}} \|^2, \quad (3)$$

where $\Theta_\mu \in \mathbb{R}^{k_1 \times c}$ and $\Theta_\sigma \in \mathbb{R}^{(k_2 - k_1) \times c}$ are the cluster association subspaces shared by target domais. Finally, we can combine $\mathcal{L}_1$ in (2) and $\mathcal{L}_2$ in (3) into a joint optimization formulation as follows:

$$\begin{aligned}\mathcal{L} =& \sum^P_{p=1} \Bigg\{ \|X^p_t - [U^p, U^p_t] \left[\begin{array}{c} \Theta^p_\mu \\ \Theta^p_t \end{array} \right] (V^p_t)^{\mathrm{T}} \|^2 \\ &+ \|X_s - [U^p, U^p_s] \left[\begin{array}{c} \Theta^p_\mu \\ \Theta^p_s \end{array} \right] Y^{\mathrm{T}}_s \|^2 \\ &+ \lambda \|X^p_t - [U^p, U^p_t] \left[\begin{array}{c} \Theta_\mu \\ \Theta_\sigma \end{array} \right] (V^p_t)^{\mathrm{T}} \|^2 \Bigg\}, \end{aligned} \quad (4)$$

where λ is the trade-off parameter weighting the contribution of target domain relatedness. The first two terms refer to the feature clusters and label propagation between source and target domains, the third term refers to the feature clusters and label updating among target domains. Overall, the proposed learning algorithm fits the multi-relevance relationship among all the domains. As we discussed in Section 2, the decomposed matrix U contains the information on hidden feature clusters, indicating the distribution of features on each hidden cluster. Therefore, the summation of each column of U has to be equal to one. The label matrix V indicates the label distribution of each instance. Thus, the summation of each row of V has to be equal to one. Considering these constraints, we obtain the final optimization objective function of the proposed learning algorithm:

$$\begin{aligned} \min_{\Omega \geq 0} \quad & \mathcal{L} \\ \text{s.t.} \quad & \sum^M_{i=1} (U^p_t)_{(ij)} = 1, \sum^M_{i=1} (U^p_s)_{(ij)} = 1, \\ & \sum^M_{i=1} U^p_{(ij)} = 1, \quad \sum^c_{j=1} (V^p_t)_{(ij)} = 1, \end{aligned} \quad (5)$$

where $\Omega = \{U^p, U^p_t, U^p_s, V^p_t, \Theta^p_\mu, \Theta^p_t, \Theta^p_s, \Theta_\mu, \Theta_\sigma\}$ is the parameter set. Since the objective function in (5) is nonconvex, it is intractable to obtain the global optimal solution. Therefore, we develop an alternating algorithm following the theory of constrained optimization (Boyd and Vandenberghe 2004). Specifically, we optimize one variable while fixing the rest variables. The procedure repeats until convergence.

We first show the updating rules of matrices U^p_t, U^p_s, U^p, and V^p_t as follows:

$$\begin{aligned} U^p_t &= U^p_t \cdot \sqrt{\frac{X^p_t V^p_t \left(\Theta^p_t\right)^{\mathrm{T}} + \lambda X^p_t V^p_t \Theta^{\mathrm{T}}_\sigma}{F_1 V^p_t \left(\Theta^p_t\right)^{\mathrm{T}} + \lambda F_3 V^p_t \Theta^{\mathrm{T}}_\sigma}}, \\ U^p_s &= U^p_s \cdot \sqrt{\frac{X_s Y_s \left(\Theta^p_s\right)^{\mathrm{T}}}{F_2 Y_s \left(\Theta^p_s\right)^{\mathrm{T}}}}, \\ U^p &= U^p \cdot \sqrt{\frac{X^p_t V^p_t \left(\Theta^p_\mu\right)^{\mathrm{T}} + X_s Y_s \left(\Theta^p_\mu\right)^{\mathrm{T}} + \lambda X^p_t V^p_t \left(\Theta_\mu\right)^{\mathrm{T}}}{F_1 V^p_t \left(\Theta^p_\mu\right)^{\mathrm{T}} + F_2 Y_s \left(\Theta^p_\mu\right)^{\mathrm{T}} + \lambda F_3 V^p_t \left(\Theta_\mu\right)^{\mathrm{T}}}}, \\ V^p_t &= V^p_t \cdot \sqrt{\frac{\left(X^p_t\right)^{\mathrm{T}} \left(U^p \Theta^p_\mu + U^p_t \Theta^p_t\right) + \left(X^p_t\right)^{\mathrm{T}} \left(U^p \Theta_\mu + U^p_t \Theta_\sigma\right)}{F^{\mathrm{T}}_1 \left(U^p \Theta^p_\mu + U^p_t \Theta^p_t\right) + \lambda F^{\mathrm{T}}_3 \left(U^p \Theta_\mu + U^p_t \Theta_\sigma\right)}}, \end{aligned} \quad (6)$$

where $F_1 = U^p \Theta^p_\mu (V^p_t)^{\mathrm{T}} + U^p_t \Theta^p_t (V^p_t)^{\mathrm{T}}$, $F_2 = U^p \Theta^p_\mu Y^{\mathrm{T}}_s + U^p_s \Theta^p_s Y^{\mathrm{T}}_s$, $F_3 = U^p \Theta_\mu (V^p_t)^{\mathrm{T}} + U^p_t \Theta_\sigma (V^p_t)^{\mathrm{T}}$, and $\cdot$ denotes matrix Hadamard product. From (5), after the matrices are updated, the constrained matrices have to be normalized as:

$$(U_t^p)_{(ij)} = \frac{(U_t^p)_{(ij)}}{\sum\limits_{i=1}^{M} (U_t^p)_{(ij)}}, \quad (U_s^p)_{(ij)} = \frac{(U_s^p)_{(ij)}}{\sum\limits_{i=1}^{M} (U_s^p)_{(ij)}},$$

$$U_{(ij)}^p = \frac{U_{(ij)}^p}{\sum\limits_{i=1}^{M} U_{(ij)}^p}, \qquad (V_t^p)_{(ij)} = \frac{(V_t^p)_{(ij)}}{\sum\limits_{j=1}^{c} (V_t^p)_{(ij)}}. \tag{7}$$

Similarly, the updating rules for other submatrices are:

$$\Theta_\mu^p = \Theta_\mu^p \cdot \sqrt{\frac{(U^p)^{\mathrm{T}} X_t^p V_t^p + (U^p)^{\mathrm{T}} X_s Y_s}{(U^p)^{\mathrm{T}} F_1 V_t^p + (U^p)^{\mathrm{T}} F_2 Y_s}},$$

$$\Theta_t^p = \Theta_t^p \cdot \sqrt{\frac{(U_t^p)^{\mathrm{T}} X_t^p V_t^p}{(U_t^p)^{\mathrm{T}} F_1 V_t^p}},$$

$$\Theta_s^p = \Theta_s^p \cdot \sqrt{\frac{(U_s^p)^{\mathrm{T}} X_s Y_s}{(U_s^p)^{\mathrm{T}} F_2 Y_s}},$$

$$\Theta_\mu = \Theta_\mu \cdot \sqrt{\frac{\sum\limits_{p=1}^{P} (U^p)^{\mathrm{T}} X_t^p V_t^p}{\sum\limits_{p=1}^{P} (U^p)^{\mathrm{T}} F_3 V_t^p}},$$

$$\Theta_\sigma = \Theta_\sigma \cdot \sqrt{\frac{\sum\limits_{p=1}^{P} (U_t^p)^{\mathrm{T}} X_t^p V_t^p}{\sum\limits_{p=1}^{P} (U_t^p)^{\mathrm{T}} F_3 V_t^p}}. \tag{8}$$

These lead to the procedure of the proposed MRTL algorithm in Algorithm 1. Moreover, as shown in (6), U^p and U_t^p are constrained by X_s, Y_s, and X_t^p. Θ_μ and Θ_σ are constrained by all the target feature matrices $\{X_t^p\}_{p=1}^P$. Therefore, the updating rule of V_t^p is constrained by X_s, Y_s, and $\{X_t^p\}_{p=1}^P$. That is, the information in the source domain and other target domains can be transferred to the p-th target.

Theoretical Analysis

This section aims to analyze the convergence property of the proposed algorithm. Without loss of generality, we formulate the detailed optimization updating of parameter U_t^p. The Lagrange function with constraint $U_t^p \geq 0$ is given by:

$$\mathcal{L} = \sum_{p=1}^{P} \mathrm{tr}\Big[(X_t^p)^{\mathrm{T}} X_t^p - 2 (X_t^p)^{\mathrm{T}} F_1 + F_1^{\mathrm{T}} F_1 + X_s^{\mathrm{T}} X_s - 2 X_s^{\mathrm{T}} F_2$$
$$+ F_2^{\mathrm{T}} F_2 + \lambda (X_t^p)^{\mathrm{T}} X_t^p - 2\lambda (X_t^p)^{\mathrm{T}} F_3 + \lambda F_3^{\mathrm{T}} F_3 \Big] \tag{9}$$
$$+ \sum_{p=1}^{P} \mathrm{tr}\left[\mathbf{\Lambda} \left((U_t^p)^{\mathrm{T}} \mathbf{1}_M - \mathbf{1}_{(k_2-k_1)} \right) \left((U_t^p)^{\mathrm{T}} \mathbf{1}_M - \mathbf{1}_{(k_2-k_1)} \right)^{\mathrm{T}} \right],$$

where $\mathbf{\Lambda} \in \mathbb{R}^{(k_2-k_1)\times(k_2-k_1)}$ is a diagonal matrix of Lagrange multiplier, $\mathbf{1}_M$ and $\mathbf{1}_{(k_2-k_1)}$ are all-ones vectors with dimension M and (k_2-k_1) respectively. Using the Karush-Kuhn-Tucker (KKT) complementarity condition, we have:

$$\frac{\partial \mathcal{L}}{\partial U_t^p} \cdot U_t^p = \Big(-2 X_t^p V_t^p (\Theta_t^p)^{\mathrm{T}} + 2 F_1 V_t^p (\Theta_t^p)^{\mathrm{T}} - 2\lambda X_t^p V_t^p \Theta_\sigma^{\mathrm{T}}$$
$$+ 2\lambda F_3 V_t^p \Theta_\sigma^{\mathrm{T}} + 2\mathbf{\Lambda} (U_t^p)^{\mathrm{T}} \mathbf{1}_M \mathbf{1}_M^{\mathrm{T}} - 2\mathbf{\Lambda} \mathbf{1}_{(k_2-k_1)} \mathbf{1}_M^{\mathrm{T}} \Big) \cdot U_t^p$$
$$= 0. \tag{10}$$

Algorithm 1: MRTL: Multi-Relevance Transfer Learning

```
1  Input: $\{X_s, Y_s\}$ from source domain, $\{X_t^p\}_{p=1}^P$ from target
   domains, number of target domains $P$, trade-off parameter $\lambda$,
   common feature clusters $k_1$, total feature clusters $k_2$, number
   of iterations $maxiter$.
2  Initialization: nomalize $X_s$ and $\{X_t^p\}_{p=1}^P$, initialize
   $\{V_t^p\}_{p=1}^P$ by logistic regression trained on source domain data
   $\{X_s, Y_s\}$.
3  iteration index $iter \leftarrow 1$.
4  while $iter < maxiter$ do
5      for $p = 1$ to $P$ do
6          update the submatrices $U_t^p, U_s^p, U^p, \Theta_\mu^p, \Theta_t^p, \Theta_s^p$, and
           label matrix $V_t^p$ according to the updating rules given
           in (6) and (8).
7          normalize the submatrices $U_t^p, U_s^p, U^p$ and label
           matrix $V_t^p$ according to the normalization rules given
           in (7).
8      update the submatrices $\Theta_\mu$ and $\Theta_\sigma$ according to the
       updating rules in (8).
9      compute objective value $\mathcal{L}^{iter}$.
10     $iter = iter + 1$.
11 Output: the predicted results $\{V_t^p\}_{p=1}^P$
```

Lemma 1. *Using the updating rule in (11) and normalization rules in (7), the loss function in (9) will monotonously decrease until convergence.*

$$U_t^p = U_t^p \cdot \sqrt{\frac{X_t^p V_t^p (\Theta_t^p)^{\mathrm{T}} + \lambda X_t^p V_t^p \Theta_\sigma^{\mathrm{T}} + \mathbf{\Lambda} \boldsymbol{1}_{(k_2-k_1)} \boldsymbol{1}_M^{\mathrm{T}}}{F_1 V_t^p (\Theta_t^p)^{\mathrm{T}} + \lambda F_3 V_t^p \Theta_\sigma^{\mathrm{T}} + \mathbf{\Lambda} (U_t^p)^{\mathrm{T}} \mathbf{1}_M \mathbf{1}_M^{\mathrm{T}}}}, \tag{11}$$

We use the auxiliary function approach (Lee and Seung 2000) to prove Lemma 1.

Lemma 2. *(Lee and Seung 2000) A funtion $G(Y, \widetilde{Y})$ is an auxiliary function for $\mathcal{T}(Y)$ if the conditions $G(Y, \widetilde{Y}) \geq \mathcal{T}(Y)$ and $G(Y, Y) = \mathcal{T}(Y)$ are satisfied for any Y, $\widetilde{Y}$. If G is an auxiliary function for $\mathcal{T}$, then $\mathcal{T}$ is non-increasing under the update*

$$Y^{(t+1)} = \arg\min_{Y} G\left(Y, Y^{(t)}\right). \tag{12}$$

Theorem 1. *Let $\mathcal{T}(U_t^p)$ denote the sum of all terms that contain U_t^p in the loss function $\mathcal{L}$ in (9) . Then the following*

$$G\left(U_t^p, \widetilde{U}_t^p\right) = -2 \sum_{ij} \left(X_t^p V_t^p (\Theta_t^p)^{\mathrm{T}} + \lambda X_t^p V_t^p \Theta_\sigma^{\mathrm{T}} \Lambda \mathbf{1}_{(k_2-k_1)} \mathbf{1}_M^{\mathrm{T}} \right)_{ij}$$
$$\left(\widetilde{U}_t^p\right)_{ij} \left(1 + \log \frac{(U_t^p)_{ij}}{(\widetilde{U}_t^p)_{ij}} \right) + \sum_{ij} \Big(F_1 V_t^p (\Theta_t^p)^{\mathrm{T}}$$
$$+ \lambda F_3 V_t^p \Theta_\sigma^{\mathrm{T}} + \Lambda (U_t^p)^{\mathrm{T}} \mathbf{1}_M \mathbf{1}_M^{\mathrm{T}} \Big)_{ij} \frac{(U_t^p)_{ij}^2}{(\widetilde{U}_t^p)_{ij}} \tag{13}$$

is an auxiliary function for $\mathcal{T}(U_t^p)$ and is a convex function in U_t^p and has a global minimum.

Theorem 1 can be proved similarly in (Ding *et al.* 2006). We omit the details here due to limited space. Based on Theorem 1, $G(U_t^p, \widetilde{U}_t^p)$ can be minimized with respect to U_t^p

and $\widetilde{U}_t^p$ fixed. Setting $\partial G(U_t^p, \widetilde{U}_t^p)/\partial U_t^p = 0$ leads to the updating rule in (11). Then Lemma 1 holds. The variable Λ in (11) still needs to be calculated. In (5), Λ is used to satisfy the condition that the summation of each column of U_t^p is 1. We use the normalization method (7) which satisfies this condition regardless of Λ. Then, $\mathbf{\Lambda} \left(U_t^p\right)^{\mathrm{T}} \mathbf{1}_M \mathbf{1}_M^{\mathrm{T}}$ is equal to $\mathbf{\Lambda 1}_{(k_2-k_1)} \mathbf{1}_M^{\mathrm{T}}$. Hence, (6) is an approximation to (11).

Theorem 2. *Using Algorithm 1 to update U_t^p, $\mathcal{T}\left(U_t^p\right)$ will monotonically decreases.*

Proof. By Lemma 2 and Theorem 1, we have $\mathcal{T}\left((U_t^p)^0\right) = G\left((U_t^p)^0, (U_t^p)^0\right) \geq G\left((U_t^p)^1, (U_t^p)^0\right) \geq G\left((U_t^p)^1, (U_t^p)^1\right) = \mathcal{T}\left((U_t^p)^1\right) \geq ...$ Therefore $\mathcal{T}\left(U_t^p\right)$ is monotonically decreasing. □

Theorem 2 also hold water with respect to the other variables. Since the objective function $\mathcal{L}$ is obviously lower bounded by 0, Algorithm 1 is guaranteed to converge.

Experiments

Experiments are tested on two benchmark data sets: 20-Newsgroups and Email Spam data sets, which are widely adopted for transfer learning evaluation.

20-Newsgroups The 20 newsgroups dataset[1] contains 18,774 documents, and has a hierarchical structure with 6 main categories and 20 subcategories. Following (Duan *et al.* 2012), we choose the instances from three main categories *comp*, *rec*, *sci*, with at least four subcategories to generate three settings to evaluate our proposed algorithms. For each setting, we choose one main category as the positive class and use another one as the negative class, and employ all the labeled instances from two subcategories to construct one domain. In the experiments, we construct one source domain and three target domains (see Table 1 for details).

Email Spam The email spam dataset[2] contains 4000 publicly available labeled emails as well as three email sets (each contains 2500 emails) annotated by three different users. Therefore, the distributions of the publicly available email set and three user-annotated email sets differ from each other. For each set, a half of the emails are non-spam (labeled as 1) and the others are spam (labeled as -1). We consider the publicly available email set as the source domain and the three user-annotated sets as three target domains.

Experimental Setup

We compare the proposed MRTL with several state-of-the-art methods: (1) Unsupervised method Nonnegative Matrix Factorization (NMF) (Lee and Seung 2000), which is directly applied to the target domain data. (2) Supervised methods, including Logistic Regression[3] (LG) and Support Vector Machine (SVM), which are trained on the source domain data and tested on the target domain data using the implementation in LibSVM[4] with linear kernel SVM. (3) Semi-supervised learning method Transductive Support Vector Machine[5](TSVM) (Joachims 1999), which works in a transductive setting using both source and target domain data for training. (4) Multi-task learning method Multi-Task Feature Learning (MTFL) (Evgeniou and Pontil 2007). It is trained on the source domain and tested on all the target domains simultaneously. (5) The state-of-the-art transfer learning methods, including Matrix Tri-Factorization based Classification (MTrick) (Zhuang *et al.* 2010), Dual Transfer Learning (DTL) (Long *et al.* 2012) and Triplex Transfer Learning (TriTL) (Zhuang *et al.* 2014). Both DTL and TriTL can be extended to solve multiple target domains by making the source and all the target domains share the same feature cluster subspace. In the experiments, the single target ones are referred to as DTL_0 and TriTL_0, and the extension ones are DTL_1 and TriTL_1, respectively. TriTL_1 and MRTL are trained using the source domain and all target domains.

Table 1: Data sets generated from 20 Newsgroups

Data Set	Source Domain	Target Domain
comp vs. rec	comp.sys.mac.hardware rec.sport.hockey	comp.graphics rec.autos
		comp.os.ms-windows.misc rec.motorcycles
		comp.sys.ibm.pc.hardware rec.sport.baseball
rec vs. sci	rec.sport.hockey sci.space	rec.autos sci.crypt
		rec.motorcycles sci.electronics
		rec.sport.baseball sci.med
sci vs. comp	sci.space comp.sys.mac.hardware	sci.crypt comp.graphics
		sci.electronics comp.os.ms-windows.misc
		sci.med comp.sys.ibm.pc.hardware

Since model selection is still an open question in transductive transfer learning, one practical solution is to choose one existing labeled data set to make training and validation. Therefore, we select *comp vs.rec* to conduct corss-validation. The parameters of the proposed method and baselines are tuned on the data set *comp vs.rec*. Then the tuned parameters are applied to all other data sets. The parameters of MRTL include the trade-off parameter λ, the number of common feature clusters k_1 and total feature clusters k_2. In the comparison experiments (see Table 2 and 3), we set $k_2 = 50$, $k_1 = 10$, $\lambda = 10$, $maxiter = 100$.

Experimental Results and Discussion

Table 2 and 3 show the accuracy of all these algorithms on each target domain and their average. We can observe from the results that the proposed MRTL consistently outperforms the considered rivals on each data set. We can also find that the non-transfer methods NMF, LG, and SVM cannot perform well on most data sets. MTFL performs poorly because without transfer the multitask classifiers trained on the source domain cannot discriminate well on target domains. TSVM outperforms them on many data sets which verifies

[1] http://people.csail.mit.edu/jrennie/20Newsgroups/

[2] http://www.ecmlpkdd2006.org/challenge.html

[3] http://research.microsoft.com/en-us/um/people/minka/papers/logreg/

[4] http://www.csie.ntu.edu.tw/ cjlin/libsvm/

[5] http://www.cs.cornell.edu/People/tj/svm_light/

Table 2: Average Classification Accuracy (%) on 20 Newsgroups Dataset

Data set	Target	NMF	LG	SVM	TSVM	MTFL	MTrick	DTL_0	DTL_1	TriTL_0	TriTL_1	MRTL
comp vs. rec	target-1	63.46	60.60	58.35	87.89	60.04	94.17	93.20	93.51	93.82	90.60	**95.09**
	target-2	59.71	65.64	64.78	92.38	66.58	93.71	93.97	96.01	94.94	92.33	**98.26**
	target-3	58.98	92.44	92.99	95.63	93.31	97.26	97.21	**97.87**	**97.82**	97.67	**97.97**
Average		60.72	72.89	72.04	91.97	73.31	95.05	94.79	95.80	95.52	93.53	**97.11**
rec vs. sci	target-1	52.58	55.31	53.85	87.60	59.41	90.08	88.41	88.41	90.03	84.16	**90.74**
	target-2	50.28	57.16	56.45	83.76	60.58	91.35	88.77	90.95	86.80	89.58	**92.46**
	target-3	63.50	85.84	86.86	92.32	87.45	96.71	**97.37**	**97.47**	95.75	96.01	96.97
Average		55.45	66.10	65.72	87.89	69.15	92.72	91.52	92.28	90.86	89.92	**93.39**
sci vs. comp	target-1	67.79	70.34	67.79	82.34	67.13	87.29	88.21	87.95	89.84	88.82	**90.30**
	target-2	57.47	60.35	59.84	66.46	59.22	75.04	76.22	76.07	76.12	74.27	**80.07**
	target-3	53.36	81.94	81.59	91.05	79.86	98.02	**98.27**	97.41	97.81	97.20	**98.58**
Average		59.54	70.88	69.74	79.95	68.74	86.78	87.57	87.14	87.92	86.76	**89.65**

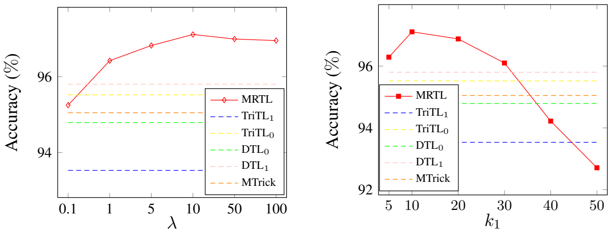
Table 3: Average Classification Accuracy (%) on Email Spam Dataset

Source	Target	NMF	LG	SVM	TSVM	MTFL	Mtrick	DTL_0	DTL_1	TriTL_0	TriTL_1	MRTL
Public Set	User 1	73.36	65.56	56.36	72.64	58.92	83.16	82.92	82.04	81.80	79.16	**83.48**
	User 2	77.80	67.28	61.32	77.92	62.84	84.36	84.04	84.52	85.16	78.76	**86.68**
	User 3	79.16	81.84	69.32	90.64	70.08	90.36	90.40	91.08	92.04	**92.68**	**92.48**
Average		76.77	71.56	62.33	80.39	63.95	85.96	85.79	86.07	86.33	83.53	**87.55**

the unlabeled data can help improve performance, but performs worse when the distribution diversity across domain is large. MTrick, DTL_0 and TriTL_0 performs better than the non-transfer methods, but have not reached the best performances becauese of the restriction that they fail to exploit the relatedness between target domains. TriTL_1 cannot simultaneously perform well on all target domains since it assumes that all the 'source-target' pairs share the same latent subspaces which would lead to insufficient transfer or overfitting on some target domains.

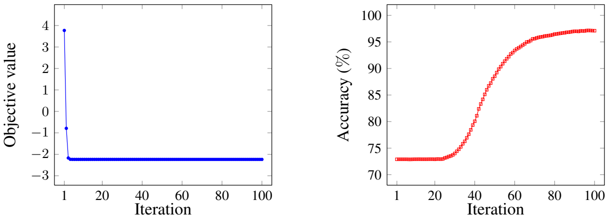
To verify that exploiting the relatedness between target domains indeed brings about effectiveness, MRTL is compared with TriTL_1, TriTL_0, DTL_0, DTL_1, and MTrick. We plot the average classification performance of MRTL with respect to λ on *comp vs. rec* data set in Figure 2(a). The average baseline results are shown as dashed lines. It can be seen that the performance of MRTL improves at first with the increasing of λ. When the parameter varies in a wide range $\lambda \in [1, 100]$, MRTL performs quite stably and consistently outperforms the baselines. It indicates that by exploiting the relatedness between target domains, MRTL achieves optimal transferability. Also, we test the model paramete k_1 varying from 5 to 50 to analyze how it affects the average classification performance. The results are shown in Figure 2(b), from which we can see that the average accuracy increases at first and then decreases, which indicates that only a part of feature clusters are shared as common, thus the partition of feature cluster subspaces is justified. The proposed method achieves better performance when k_1 is between 5 and 30. For different 'source-target' pairs, we can set different k_1 values. In this paper, we simply set them equal.

In Section 4, we have theoretically proven the convergence property of the proposed MRTL algorithm. Here we empirically check the convergence by testing it on *comp vs. rec* data set. In Figure 3(a), we show the logarithmic objective value with respect to the number of iterations. We see that after around five iterations, the objective value experiences almost no change. Similarly, we show the average classification accuracy of all target domains with respect to the number of iterations in Figure 3(b). The results show that the average accuracy of MRTL increases with more iterations and converges after 80 iterations.



(a) (b)

Figure 2: Performance of MRTL with respect to λ and k_1 on *comp vs. rec* data set.



(a) Objective values (b) Accuracy

Figure 3: Performance of MRTL with respect to iterations on *comp vs. rec* data set.

Conclusion

In this paper, we study multi-relevance transfer learning, where knowledge not only needs to be transferred from the source domain but also from all the target domains. We propose a MRTL framework to solve this problem. The framework achieves both source-target transfer and target-target transfer by sharing multiple decomposed latent subspaces. We develop an alternating scheme for optimization. Experiments on two datasets show the effectiveness of the proposed approach. The convergence property has also been theoretically and experimentally proven. In future, extending MRTL to tackle online tasks is an interesting problem.

References

John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *EMNLP*, pages 120–128. Association for Computational Linguistics, 2006.

John Blitzer, Dean Foster, and Sham Kakade. Domain adaptation with coupled subspaces. In *AISTATS*, 2011.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Olivier Chapelle, Pannagadatta Shivaswamy, Srinivas Vadrevu, Kilian Weinberger, Ya Zhang, and Belle Tseng. Multi-task learning for boosting with application to web search ranking. In *SIGKDD*, pages 1189–1198. ACM, 2010.

Rita Chattopadhyay, Qian Sun, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Multisource domain adaptation and its application to early detection of fatigue. *TKDD*, 6(4):18, 2012.

Wei Cheng, Xiang Zhang, Zhishan Guo, Yubao Wu, Patrick F Sullivan, and Wei Wang. Flexible and robust co-regularized multi-domain graph clustering. In *SIGKDD*, pages 320–328. ACM, 2013.

Chris H. Q. Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *SIGKDD, Philadelphia, PA, USA, August 20-23, 2006*, pages 126–135, 2006.

Mark Dredze, Alex Kulesza, and Koby Crammer. Multi-domain learning by confidence-weighted parameter combination. *Machine Learning*, 79(1-2):123–149, 2010.

Lixin Duan, Dong Xu, and Ivor W. Tsang. Domain adaptation from multiple sources: A domain-dependent regularization approach. *IEEE Transactions on Neural Networks and Learning Systems*, 23(3):504–518, March 2012.

Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *SIGKDD, Seattle, Washington, USA, August 22-25, 2004*, pages 109–117, 2004.

A Evgeniou and Massimiliano Pontil. Multi-task feature learning. *NIPS*, 19:41, 2007.

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

Sunil Kumar Gupta, Dinh Phung, Brett Adams, Truyen Tran, and Svetha Venkatesh. Nonnegative shared subspace learning and its application to social media retrieval. In *SIGKDD*, pages 1169–1178. ACM, 2010.

Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in NLP. In *ACL, June 23-30, 2007, Prague, Czech Republic*, 2007.

Thorsten Joachims. Transductive inference for text classification using support vector machines. In *ICML*, volume 99, pages 200–209, 1999.

Daniel D. Lee and H. Sebastian Seung. Algorithms for nonnegative matrix factorization. In *NIPS*, pages 556–562. MIT Press, 2000.

Mingsheng Long, Jianmin Wang, Guiguang Ding, Wei Cheng, Xiang Zhang, and Wei Wang. Dual transfer learning. In *SDM, California, USA, April 26-28, 2012.*, pages 540–551, 2012.

Guillaume Obozinski, Ben Taskar, and Michael Jordan. Multi-task feature selection. *Statistics Department, UC Berkeley, Tech. Rep*, 2006.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.

Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. In *IJCAI, USA, July 11-17, 2009*, pages 1187–1192, 2009.

Ben Tan, Yangqiu Song, ErHeng Zhong, and Qiang Yang. Transitive transfer learning. In *SIGKDD, Sydney, NSW, Australia, August 10-13, 2015*, pages 1155–1164, 2015.

Yu Zhang, Bin Cao, and Dit-Yan Yeung. Multi-domain collaborative filtering. *arXiv:1203.3535*, 2012.

K. Zhang, M. Gong, and B. Schölkopf. Multi-source domain adaptation: A causal view. In *AAAI*, pages 3150–3157. AAAI Press, 2015.

Fuzhen Zhuang, Ping Luo, Hui Xiong, Qing He, Yuhong Xiong, and Zhongzhi Shi. Exploiting associations between word clusters and document classes for cross-domain text categorization. In *SDM 2010, April 29 - May 1, 2010, Columbus, Ohio, USA*, pages 13–24, 2010.

Fuzhen Zhuang, Ping Luo, Changying Du, Qing He, Zhongzhi Shi, and Hui Xiong. Triplex transfer learning: exploiting both shared and distinct concepts for text classification. *Cybernetics, IEEE Transactions on*, 44(7):1191–1203, 2014.