Transfer Learning with Graph Co-Regularization

Mingsheng Long[†‡§], **Jianmin Wang**[†§], **Guiguang Ding**[†§], **Dou Shen**[♮] and **Qiang Yang**[♯]
[†]Tsinghua National Laboratory for Information Science and Technology
[‡]Department of Computer Science and Technology, Tsinghua University
[§]School of Software, Tsinghua University, Beijing 100084, China
[♮]CityGrid Media, Seattle, WA 98104, USA
[♯]Hong Kong University of Science and Technology, Hong Kong
longmingsheng@gmail.com, {jimwang, dinggg}@tsinghua.edu.cn, doushen@live.com, qyang@cse.ust.hk

Abstract

Transfer learning proves to be effective for leveraging labeled data in the source domain to build an accurate classifier in the target domain. The basic assumption behind transfer learning is that the involved domains share some common latent factors. Previous methods usually explore these latent factors by optimizing two separate objective functions, i.e., either maximizing the empirical likelihood, or preserving the geometric structure. Actually, these two objective functions are complementary to each other and optimizing them simultaneously can make the solution smoother and further improve the accuracy of the final model. In this paper, we propose a novel approach called *Graph co-regularized Transfer Learning* (GTL) for this purpose, which integrates the two objective functions seamlessly into one unified optimization problem. Thereafter, we present an iterative algorithm for the optimization problem with rigorous analysis on convergence and complexity. Our empirical study on two open data sets validates that GTL can consistently improve the classification accuracy compared to the state-of-the-art transfer learning methods.

Introduction

A striking aspect of machine learning is the ability to process a large amount of unorganized information by learning models from labeled data in a domain. Unfortunately, it is often very expensive to obtain sufficient labeled data. In the meanwhile, it is not uncommon that abundant labeled data exist in other relevant domains. However, traditional machine learning algorithms cannot leverage the labeled data from different domains since they require the same distribution across all the data. To address this situation, transfer learning (Pan and Yang 2010) has been widely studied to explore the divergent distributions among multiple data domains and train accurate classifiers in the domains of interest. It proves to be very effective in many real-life applications such as text categorization (Zhuang et al. 2011), sentiment analysis (Pan et al. 2010a), image classification (Zhu et al. 2011) and collaborative filtering (Pan et al. 2010b).

The basic assumption behind transfer learning is that there exist some common latent factors shared by all domains. These latent factors can be exploited to reduce the distribution divergence and bridge different domains. Previous methods usually uncover these latent factors by optimizing predefined objective functions, including maximizing the empirical likelihood (Dai et al. 2007; Zhuang et al. 2010; Wang et al. 2011; Zhuang et al. 2011), and preserving the intrinsic geometric structure (Ling et al. 2008; Pan et al. 2011; Wang and Mahadevan 2009; 2011). Actually, these objective functions focus on different aspects of the data and are complementary to each other to some extent (Zhu and Lafferty 2005). On one hand, each data point may be associated with some latent factors. For example, a text document can be regarded as a combination of several hidden topics. Extracting these latent factors involves maximizing the empirical likelihood of data statistics. On the other hand, from a geometric perspective, the data points may be sampled from a distribution supported by a low-dimensional manifold embedded in a high-dimensional space (Cai et al. 2009). This geometric structure, meaning close samples tend to have the same label, should be preserved when the common latent factors are discovered as the bridge for knowledge transfer. Otherwise, transfer learners based on the discovered bridge may fail to predict labels smoothly with respect to the geometric structure. However, most of the existing works in the literature consider the two objective functions separately without exploring the benefit of integrating them in a unified manner.

In this paper, we address the aforementioned issue by developing a novel approach called Graph co-regularized Transfer Learning (GTL). GTL combines the maximization of empirical likelihood and the preservation of geometric structure into one unified objective function so that it can optimize the empirical likelihood and the geometric structure simultaneously. As a result, GTL can successfully explore the common latent factors to facilitate smooth transfer learning without breaking the geometric structure.

GTL is implemented as a novel algorithm named Graph co-regularized Collective Matrix tri-Factorization (GCMF). The main idea of GCMF is illustrated in Figure 1. GCMF works for any number of domains, but for the ease of explanation, we use two domains: one source domain $\mathcal{D}_s$ and one target domain $\mathcal{D}_t$. The domain indices are $\mathcal{I} = \{s, t\}$. Each domain $\mathcal{D}_\pi$, $\pi \in \mathcal{I}$ has a feature-example matrix $\mathbf{X}_\pi$. To uncover the latent factors, we decompose $\mathbf{X}_\pi$ into a product of three nonnegative matrices, i.e., $\mathbf{X}_\pi = \mathbf{U}_\pi \mathbf{H} \mathbf{V}_\pi^{\mathrm{T}}$, where

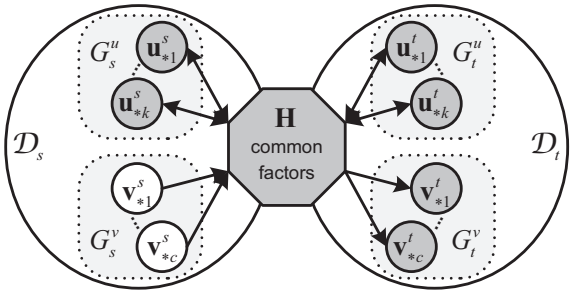


Figure 1: Illustration of graph co-regularized collective matrix tri-factorization (GCMF) algorithm for GTL. Small circles represent latent factors, among which the unfilled ones are the groundtruth example classes in the source domain.

$\mathbf{U}_\pi = [\mathbf{u}^\pi_{*1}, \ldots, \mathbf{u}^\pi_{*k}]$, $\mathbf{V}_\pi = [\mathbf{v}^\pi_{*1}, \ldots, \mathbf{v}^\pi_{*c}]$, and $\mathbf{a}_{*i}$ is the ith column of matrix $\mathbf{A}$. The matrix tri-factorization performs feature-example co-clustering by maximizing the empirical likelihood in domain $\mathcal{D}_\pi$ (Ding, Li, and Peng 2006), with $\mathbf{H}$ representing the association between feature clusters $\mathbf{U}_\pi$ and example classes $\mathbf{V}_\pi$. $\mathbf{H}$ is shown to remain stable across domains, which can therefore be treated as a shared bridge for knowledge transfer (Zhuang et al. 2011). Furthermore, we construct affinity graphs G^u_π and G^v_π to encode the geometrical information underlying the feature space and the example space in domain $\mathcal{D}_\pi$, respectively. Intuitively, similar features tend to represent the same semantic, while close examples tend to have the same label. Taking these two graphs as co-regularization for the matrix tri-factorization, the geometrical information is seamlessly incorporated into the co-clustering procedure, making the labels predicted by the learned model sufficiently smooth with respect to the intrinsic geometric structure. This will guarantee the uncovered common latent factors $\mathbf{H}$ to be a smooth bridge and facilitate more effective transfer learning.

The main contribution of this paper is the GTL framework to address transfer learning with constrained information. We specifically study the simultaneous optimization of empirical likelihood and geometric structure by formulating a matrix factorization with graph regularization. We believe that GTL can handle transfer learning when any prior knowledge is available and can be incorporated into graph regularization. Typical prior knowledge includes links in network mining, or user trust relations in collaborative filtering. It can also be easily extended to handle multiple data domains. Extensive experiments on the widely adopted data sets validate the effectiveness of our proposed approach.

Related Work

Previous works in transfer learning assume a set of common latent factors existing across the source domain and the target domain, which behave as a bridge to transfer knowledge between them (Pan and Yang 2010). These latent factors are discovered by optimizing certain predefined objective functions, such as empirical likelihood and geometric structure.

Collective Matrix Factorization (CMF) (Singh and Gordon 2008) and its tri-factorization variants have been extensively studied for transfer learning recently (Gupta et al. 2010; Long et al. 2010; Wang et al. 2011; Zhu et al. 2011; Zhuang et al. 2011; Long et al. 2012). CMF jointly factorizes multiple matrices with correspondences between rows and columns while enforcing a set of common latent factors that match rows and columns across different matrices. The common latent factors are then used as a bridge for knowledge transfer. Existing CMF-based methods mainly maximize the empirical likelihood among multiple domains (Singh and Gordon 2008). If we can explore deeper in the data beyond the empirical likelihood, we may discover more connections among the domains and build up a better transfer model.

Recently, the geometric structure has been explored for transfer learning, including Cross-Domain Spectral Classification (CDSC) (Ling et al. 2008), Transfer Component Analysis (TCA) (Pan et al. 2011) and Manifold Alignment (MA) (Wang and Mahadevan 2009; 2011). CDSC and TCA seek consistency between the in-domain supervision and the out-of-domain geometric structure via spectral learning. MA maps different domains to a new latent space, simultaneously matching corresponding examples and preserving the geometric structure of each domain. Contrary to the CMF based methods, these methods focus only on the geometric structure and do not optimize the empirical likelihood.

In this paper, we put forward an algorithm called Graph co-regularized Collective Matrix tri-Factorization (GCMF) to simultaneously maximize the empirical likelihood and preserve the geometric structure across domains so that we can combine the advantages of the aforementioned two sets of methods. In GCMF, we explore the geometrical information underlying both example space and feature space when considering the geometric structure, while CDSC, TCA and MA only handle the geometrical information underlying the example space. Beyond the geometric structure, GTL can be treated as a general framework in which we can easily incorporate various prior knowledge, such as links in network mining and user trust relations in collaborative filtering.

Graph Co-Regularized Transfer Learning

Problem Definition

We focus on *transductive transfer learning* where the source domain has abundant labeled examples while the target domain has only unlabeled data. We consider one source domain $\mathcal{D}_s$ and one target domain $\mathcal{D}_t$. The domain indices are $\mathcal{I} = \{s, t\}$. $\mathcal{D}_s$ and $\mathcal{D}_t$ share the same feature space and label space. There are m features and c classes. Let $\mathbf{X}_\pi = [\mathbf{x}^\pi_{*1}, \ldots, \mathbf{x}^\pi_{*n_\pi}] \in \mathbb{R}^{m \times n_\pi}, \pi \in \mathcal{I}$ represent the feature-example matrix of domain $\mathcal{D}_\pi$, where $\mathbf{x}^\pi_{*i}$ is the ith example in domain $\mathcal{D}_\pi$. Labels of the examples in the source domain $\mathcal{D}_\pi$ are given as $\mathbf{Y}_\pi \in \mathbb{R}^{n_\pi \times c}$, where $y^\pi_{ij} = 1$ if $\mathbf{x}^\pi_{*i}$ belongs to class j, and $y^\pi_{ij} = 0$ otherwise. Frequently used notations and descriptions are summarized in Table 1.

The goal of *Graph co-regularized Transfer Learning* (GTL) is to uncover the common latent factors underlying multiple domains as the bridge for knowledge transfer which simultaneously (1) maximizes the empirical likelihood of all domains, (2) preserves the geometric structure in each domain.

Table 1: Notations and descriptions used in this paper.

Notation	Description
$\mathcal{I}$	domain indices, $\mathcal{I} = \{s, t\}$
$\mathcal{D}_\pi$	domain π, $\pi \in \mathcal{I}$
n_π	#examples in domain $\mathcal{D}_\pi$
m	#features in the shared feature space
c	#classes in the shared label space
k, p	#feature clusters, #nearest neighbors
λ, γ	regularization parameters
$\mathbf{X}_\pi$	$m \times n_\pi$ data matrix of domain $\mathcal{D}_\pi$
$\mathbf{Y}_\pi$	$n_\pi \times c$ label matrix of domain $\mathcal{D}_\pi$
$\mathbf{U}_\pi$	k feature clusters in domain $\mathcal{D}_\pi$
$\mathbf{V}_\pi$	c example classes in domain $\mathcal{D}_\pi$
$\mathbf{H}_\pi, \mathbf{H}$	association between $\mathbf{U}_\pi$ and $\mathbf{V}_\pi$
G^u_π, G^v_π	feature graph, example graph in domain $\mathcal{D}_\pi$
$\mathbf{a}_{*i}$, $\mathbf{a}_{i*}$	ith column of matrix $\mathbf{A}$, ith row of matrix $\mathbf{A}$

The Proposed Approach

To achieve the goal of GTL, we propose an algorithm called *Graph co-regularized Collective Matrix tri-Factorization* (GCMF) which can seamlessly optimize both the empirical likelihood and the geometric structure for effective transfer learning. An illustration of GCMF is given in Figure 1.

Collective Matrix Tri-Factorization We first discover the latent factors in each domain $\mathcal{D}_\pi$. These latent factors can be uncovered using nonnegative matrix tri-factorization (NMTF) (Ding et al. 2006) which performs co-clustering on the feature-example matrix $\mathbf{X}_\pi, \forall \pi \in \mathcal{I}$,

$$\min_{\mathbf{U}_\pi, \mathbf{H}_\pi, \mathbf{V}_\pi \geq \mathbf{0}} \mathcal{L}_\pi = \left\| \mathbf{X}_\pi - \mathbf{U}_\pi \mathbf{H}_\pi \mathbf{V}_\pi^{\mathrm{T}} \right\|^2 \tag{1}$$

where $\|\mathbf{A}\|$ is the Frobenius norm of matrix $\mathbf{A}$. $\mathbf{U}_\pi = [\mathbf{u}^\pi_{*1}, \ldots, \mathbf{u}^\pi_{*k}] \in \mathbb{R}^{m \times k}$, each $\mathbf{u}^\pi_{*i}$ represents a semantic concept, i.e., a feature cluster. $\mathbf{V}_\pi = [\mathbf{v}^\pi_{*1}, \ldots, \mathbf{v}^\pi_{*c}] \in \mathbb{R}^{n_\pi \times c}$, each $\mathbf{v}^\pi_{*i}$ represents an example class. $\mathbf{U}_\pi$ and $\mathbf{V}_\pi$ are the co-clustering results on features and examples respectively. $\mathbf{H}_\pi \in \mathbb{R}^{k \times c}$ is the association between feature clusters $\mathbf{U}_\pi$ and example classes $\mathbf{V}_\pi$, which has been shown to remain more stable underlying different domains than $\mathbf{U}_\pi$, $\mathbf{V}_\pi$ (Zhuang et al. 2011). Therefore, we can assume $\mathbf{H}_\pi = \mathbf{H}$ for each domain $\mathcal{D}_\pi$. This leads to the following *collective matrix tri-factorization* formulation

$$\min_{\mathbf{U}_\pi, \mathbf{H}, \mathbf{V}_\pi \geq \mathbf{0}} \mathcal{L} = \sum_{\pi \in \mathcal{I}} \left\| \mathbf{X}_\pi - \mathbf{U}_\pi \mathbf{H} \mathbf{V}_\pi^{\mathrm{T}} \right\|^2 \tag{2}$$

The common latent factors $\mathbf{H}$ are uncovered as a stable bridge for knowledge transfer, with the enforcement of the supervised information in the source domain, i.e., enforcing $\mathbf{V}_s \equiv \mathbf{Y}_s$. Through the bridge, knowledge from source domain labels $\mathbf{Y}_s$ is transferred to the target domain samples $\mathbf{V}_t$. This procedure corresponds to maximizing the empirical likelihood of multiple domains (Ding, Li, and Peng 2006).

Graph Co-Regularization From a geometric perspective, the data points may be sampled from a distribution supported by a low-dimensional manifold embedded in a high-dimensional space (Cai et al. 2009). So we hope that the geometric structure can be preserved during the transfer procedure to avoid violating the intrinsic data distributions. By the *manifold assumption* (Belkin and Niyogi 2001), if two examples $\mathbf{x}^\pi_{*i}, \mathbf{x}^\pi_{*j}$ in domain $\mathcal{D}_\pi$ are close in the intrinsic geometry of the data distribution, then their labels $\mathbf{v}^\pi_{i*}$ and $\mathbf{v}^\pi_{j*}$ should also be close. The geometric structure can be modeled by a nearest neighbor graph in the example space. Consider an *example graph* G^v_π with n_π vertices each representing an example in domain $\mathcal{D}_\pi$. Define the affinity matrix

$$(\mathbf{W}^v_\pi)_{ij} = \begin{cases} \cos\left(\mathbf{x}^\pi_{*i}, \mathbf{x}^\pi_{*j}\right), \text{if } \mathbf{x}^\pi_{*i} \in \mathcal{N}_p\left(\mathbf{x}^\pi_{*j}\right) \text{ or } \mathbf{x}^\pi_{*j} \in \mathcal{N}_p\left(\mathbf{x}^\pi_{*i}\right) \\ 0, \text{otherwise} \end{cases} \tag{3}$$

where $\mathcal{N}_p\left(\mathbf{x}^\pi_{*i}\right)$ denotes the set of p nearest neighbors of example $\mathbf{x}^\pi_{*i}$. Let $\mathbf{D}^v_\pi = \mathrm{diag}\left(\sum_i (\mathbf{W}^v_\pi)_{ij}\right)$. Preserving the geometric structure in domain $\mathcal{D}_\pi$ is reduced to minimizing the example graph regularizer

$$\begin{aligned} \mathcal{R}^v_\pi =& \frac{1}{2} \sum_{ij} \left\| \mathbf{v}^\pi_{i*} - \mathbf{v}^\pi_{j*} \right\|^2 (\mathbf{W}^v_\pi)_{ij} \\ =& \sum_i \mathbf{v}^\pi_{i*} (\mathbf{v}^\pi_{i*})^{\mathrm{T}} (\mathbf{W}^v_\pi)_{ii} - \sum_{ij} \mathbf{v}^\pi_{i*} (\mathbf{v}^\pi_{j*})^{\mathrm{T}} (\mathbf{W}^v_\pi)_{ij} \\ =& \mathrm{tr}\left(\mathbf{V}^{\mathrm{T}}_\pi \left(\mathbf{D}^v_\pi - \mathbf{W}^v_\pi \right) \mathbf{V}_\pi \right) \end{aligned} \tag{4}$$

Furthermore, from the duality property between features and examples, the features are also sampled from a distribution supported by another low-dimensional manifold (Gu and Zhou 2009). Thus we construct a *feature graph* G^u_π with m vertices each representing a feature in domain $\mathcal{D}_\pi$ as

$$(\mathbf{W}^u_\pi)_{ij} = \begin{cases} \cos\left(\mathbf{x}^\pi_{i*}, \mathbf{x}^\pi_{j*}\right), \text{if } \mathbf{x}^\pi_{i*} \in \mathcal{N}_p\left(\mathbf{x}^\pi_{j*}\right) \text{ or } \mathbf{x}^\pi_{j*} \in \mathcal{N}_p\left(\mathbf{x}^\pi_{i*}\right) \\ 0, \text{otherwise} \end{cases} \tag{5}$$

where $\mathcal{N}_p\left(\mathbf{x}^\pi_{i*}\right)$ denotes the set of p nearest neighbors of feature $\mathbf{x}^\pi_{i*}$. Let $\mathbf{D}^u_\pi = \mathrm{diag}\left(\sum_i (\mathbf{W}^u_\pi)_{ij}\right)$. Preserving the geometric structure in domain $\mathcal{D}_\pi$ further includes minimizing the feature graph regularizer

$$\mathcal{R}^u_\pi = \frac{1}{2} \sum_{ij} \left\| \mathbf{u}^\pi_{i*} - \mathbf{u}^\pi_{j*} \right\|^2 (\mathbf{W}^u_\pi)_{ij} = \mathrm{tr}\left(\mathbf{U}^{\mathrm{T}}_\pi \left(\mathbf{D}^u_\pi - \mathbf{W}^u_\pi \right) \mathbf{U}_\pi \right) \tag{6}$$

Optimization Clearly, the graph regularizers defined in Equations (4) and (6) can preserve the geometric structure for the co-clustering on examples and features. Therefore, we can treat them as *graph co-regularization* and incorporate them into Equation (2). This allows us to reach the optimization problem of *Graph co-regularized Collective Matrix tri-Factorization* (GCMF) as defined in Equation (7)

$$\begin{aligned} &\min_{\mathbf{U}_\pi, \mathbf{H}, \mathbf{V}_\pi \geq \mathbf{0}} \mathcal{O} = \mathcal{L} + \sum_{\pi \in \mathcal{I}} \left(\lambda \mathcal{R}^u_\pi + \gamma \mathcal{R}^v_\pi \right) \\ &s.t. \quad \mathbf{U}^{\mathrm{T}}_\pi \mathbf{1}_m = \mathbf{1}_k, \mathbf{V}^{\mathrm{T}}_\pi \mathbf{1}_{n_\pi} = \mathbf{1}_c, \forall \pi \in \mathcal{I} \end{aligned} \tag{7}$$

where λ, γ are regularization parameters. The ℓ_1 normalization constraints on each column of $\mathbf{U}_\pi$ and $\mathbf{V}_\pi$ are used to make the optimization well-defined (Gu, Ding, and Han 2011). With the optimization results, the label for any example $\mathbf{x}^\pi_{*i}$ in the target domain $\mathcal{D}_\pi$ can be easily inferred by

$$f\left(\mathbf{x}^\pi_{*i}\right) = \arg\max_j \left(\mathbf{V}_\pi\right)_{ij}$$

In optimization (7), the discovered common latent factors $\mathbf{H}$ are subject to maximizing the empirical likelihood and preserving the geometric structure simultaneously. Therefore, their smoothness is enhanced during the transfer procedure.

GCMF can be easily extended to handle multiple domains by expanding $\mathcal{I}$ to include more domains and exploring their common underlying structure collectively. It can also handle transfer learning when some prior knowledge is available and can be incorporated into graph regularization, by following the graph construction steps as defined in Equations (3) and (5).

Learning Algorithm

The optimal solution to the GCMF optimization problem in Equation (7) is achieved through the following theorem. More detailed analysis is presented in the next subsection.

Theorem 1 *Updating $\mathbf{U}_\pi$, $\mathbf{V}_\pi$, $\mathbf{H}$ with Equations* (8)∼(10) *for each domain $\mathcal{D}_\pi, \forall \pi \in \mathcal{I}$ will monotonically decrease the objective function in Equation* (7) *until convergence.*

$$\mathbf{U}_\pi \leftarrow \mathbf{U}_\pi \circ \sqrt{\frac{[\mathbf{X}_\pi \mathbf{V}_\pi \mathbf{H}^{\mathrm{T}} + \lambda \mathbf{W}_\pi^u \mathbf{U}_\pi]}{\left[\mathbf{U}_\pi \mathbf{H} \mathbf{V}_\pi^{\mathrm{T}} \mathbf{V}_\pi \mathbf{H}^{\mathrm{T}} + \lambda \mathbf{D}_\pi^u \mathbf{U}_\pi\right]}} \tag{8}$$

$$\mathbf{V}_\pi \leftarrow \mathbf{V}_\pi \circ \sqrt{\frac{[\mathbf{X}_\pi^{\mathrm{T}} \mathbf{U}_\pi \mathbf{H} + \gamma \mathbf{W}_\pi^v \mathbf{V}_\pi]}{[\mathbf{V}_\pi \mathbf{H}^{\mathrm{T}} \mathbf{U}_\pi^{\mathrm{T}} \mathbf{U}_\pi \mathbf{H} + \gamma \mathbf{D}_\pi^v \mathbf{V}_\pi]}} \tag{9}$$

$$\mathbf{H} \leftarrow \mathbf{H} \circ \sqrt{\frac{\left[\sum_{\pi \in \mathcal{I}} \mathbf{U}_\pi^{\mathrm{T}} \mathbf{X}_\pi \mathbf{V}_\pi\right]}{\left[\sum_{\pi \in \mathcal{I}} \mathbf{U}_\pi^{\mathrm{T}} \mathbf{U}_\pi \mathbf{H} \mathbf{V}_\pi^{\mathrm{T}} \mathbf{V}_\pi\right]}} \tag{10}$$

where $\circ$ denotes element-wise product, $\frac{[\cdot]}{[\cdot]}$ denotes element-wise division, $\sqrt{\cdot}$ denotes element-wise square root.

Algorithm 1: GCMF: Graph Co-Regularized Collective Matrix Tri-Factorization for Transfer Learning

Input: data sets $\{\mathbf{X}_\pi\}_{\pi \in \mathcal{I}}$, $\mathbf{Y}_s$, parameters k, p, λ, γ.
Output: classification results in the target domain $\mathbf{V}_t$.
begin
  construct graphs G_π^v, G_π^u by Equations (3) and (5).
  normalize data sets by $\mathbf{X}_\pi \leftarrow \mathbf{X}_\pi / \|\mathbf{X}_\pi\|, \forall \pi \in \mathcal{I}$.
  initialize $\{\mathbf{U}_\pi\}_{\pi \in \mathcal{I}}$, $\mathbf{H}$ by random positives, $\mathbf{V}_s$ by $\mathbf{Y}_s$, $\mathbf{V}_t$ by logistic regression trained on $\{\mathbf{X}_s, \mathbf{Y}_s\}$.
  for $iter \leftarrow 1$ **to** $maxIter$ **do**
    foreach $\pi \in \mathcal{I}$ **do**
      update $\mathbf{U}_\pi, \mathbf{V}_\pi, \mathbf{H}$ by Equations (8)∼(10) fixing $\mathbf{V}_s \equiv \mathbf{Y}_s$.
      normalize each column of $\mathbf{U}_\pi, \mathbf{V}_\pi$ by ℓ_1 norm.
      compute objective $\mathcal{O}^{iter}$ by Equation (7).
end

With Theorem 1, the learning algorithm for the GCMF optimization is summarized in Algorithm 1. To make the algorithm converge faster, we train a logistic regression classifier over the source domain and use it to classify the examples in the target domain so that we can initialize the estimated labels of the target domain better than random assignments. We keep the labels of the source domain data fixed, i.e., $\mathbf{V}_s \equiv \mathbf{Y}_s$, instead of updating them during iterations. The time complexity of Algorithm 1 is $\mathcal{O}\left(\sum_{\pi \in \mathcal{I}} \left(maxIter \cdot kmn_\pi + mn_\pi^2 + m^2 n_\pi\right)\right)$ on $|\mathcal{I}|$ domains, which is the sum of NMTF cost plus the nearest neighbor graph construction cost in each domain.

Theoretical Analysis

Derivation We derive the solution to the optimization problem in Equation (7) following the theory of constrained optimization (Boyd and Vandenberghe 2004). Specifically, we optimize one variable and derive its updating rule while fixing the rest variables. The procedure repeats until convergence. We formulate a Lagrange function for the optimization with nonnegative and ℓ_1 normalization constraints as

$$\begin{aligned} L = \mathcal{O} &+ \sum\nolimits_{\pi \in \mathcal{I}} \mathrm{tr}\left(\mathbf{\Lambda}_\pi \left(\mathbf{U}_\pi^{\mathrm{T}} \mathbf{1}_m - \mathbf{1}_k\right) \left(\mathbf{U}_\pi^{\mathrm{T}} \mathbf{1}_m - \mathbf{1}_k\right)^{\mathrm{T}}\right) \\ &+ \sum\nolimits_{\pi \in \mathcal{I}} \mathrm{tr}\left(\mathbf{\Gamma}_\pi \left(\mathbf{V}_\pi^{\mathrm{T}} \mathbf{1}_{n_\pi} - \mathbf{1}_c\right) \left(\mathbf{V}_\pi^{\mathrm{T}} \mathbf{1}_{n_\pi} - \mathbf{1}_c\right)^{\mathrm{T}}\right) \end{aligned}$$

where $\mathbf{\Lambda}_\pi \in \mathbb{R}^{k \times k}$, $\mathbf{\Gamma}_\pi \in \mathbb{R}^{c \times c}$ are the Lagrange multipliers for the two constraints, $\mathbf{1}_m$ is the vector of ones. We then derive the updating rule for $\mathbf{U}_\pi$ using the Karush-Kuhn-Tucker (KKT) complementarity condition (Boyd and Vandenberghe 2004) for the constraints on $\mathbf{U}_\pi$ we have

$$\begin{aligned} &\frac{1}{2} \nabla_{\mathbf{U}_\pi} L \circ \mathbf{U}_\pi = \left(\mathbf{V}_\pi \mathbf{H}^{\mathrm{T}} \mathbf{U}_\pi^{\mathrm{T}} \mathbf{U}_\pi \mathbf{H} - \mathbf{X}_\pi \mathbf{V}_\pi \mathbf{H}^{\mathrm{T}}\right) \circ \mathbf{U}_\pi \\ &+ \left(\lambda \mathbf{D}_\pi^u \mathbf{U}_\pi - \lambda \mathbf{W}_\pi^u \mathbf{U}_\pi + \mathbf{1}_m \mathbf{1}_m^{\mathrm{T}} \mathbf{U}_\pi \mathbf{\Lambda}_\pi - \mathbf{1}_m \mathbf{1}_k^{\mathrm{T}} \mathbf{\Lambda}_\pi\right) \circ \mathbf{U}_\pi = \mathbf{0} \end{aligned}$$

The KKT condition leads to the following updating formula

$$\mathbf{U}_\pi \leftarrow \mathbf{U}_\pi \circ \sqrt{\frac{[\mathbf{X}_\pi \mathbf{V}_\pi \mathbf{H}^{\mathrm{T}} + \lambda \mathbf{W}_\pi^u \mathbf{U}_\pi + \mathbf{1}_m \mathbf{1}_k^{\mathrm{T}} \mathbf{\Lambda}_\pi]}{\left[\mathbf{U}_\pi \mathbf{H} \mathbf{V}_\pi^{\mathrm{T}} \mathbf{V}_\pi \mathbf{H}^{\mathrm{T}} + \lambda \mathbf{D}_\pi^u \mathbf{U}_\pi + \mathbf{1}_m \mathbf{1}_m^{\mathrm{T}} \mathbf{U}_\pi \mathbf{\Lambda}_\pi\right]}}$$

We avoid computing $\mathbf{\Lambda}_\pi$ by using an iterative normalization technique (Zhuang et al. 2010). For each iteration, we normalize each column of $\mathbf{U}_\pi$ in Algorithm 1 so that $\mathbf{U}_\pi^{\mathrm{T}} \mathbf{1}_m = \mathbf{1}_k$. After that, we obtain two equal terms $\mathbf{1}_m \mathbf{1}_m^{\mathrm{T}} \mathbf{U}_\pi \mathbf{\Lambda}_\pi = \mathbf{1}_m \mathbf{1}_k^{\mathrm{T}} \mathbf{\Lambda}_\pi$ that depend only on $\mathbf{\Lambda}_\pi$. They can be omitted from the updating formula without impacting the convergence. This leads to the updating rule for $\mathbf{U}_\pi$ in Equation (8).

Convergence We use the auxiliary function approach (Lee and Seung 2000) to prove the convergence property of Theorem 1 and Algorithm 1. Firstly, the definitions and properties of the auxiliary function are introduced as follows.

Definition 1 *(Lee and Seung 2000) $A(\mathbf{Z}, \tilde{\mathbf{Z}})$ is an auxiliary function for $L(\mathbf{Z})$ if the conditions*

$$A(\mathbf{Z}, \tilde{\mathbf{Z}}) \geq L(\mathbf{Z}) \text{ and } A(\mathbf{Z}, \mathbf{Z}) = L(\mathbf{Z})$$

are satisfied for any given $\mathbf{Z}$, $\tilde{\mathbf{Z}}$.

Lemma 1 *(Lee and Seung 2000) If A is an auxiliary function for L, then L is non-increasing under the update*

$$\mathbf{Z}^{(t+1)} = \arg\min_{\mathbf{Z}} A(\mathbf{Z}, \mathbf{Z}^{(t)})$$

Theorem 2 *Let $L(\mathbf{U}_\pi)$ denote the sum of all terms in L that contain $\mathbf{U}_\pi$. The following function*

$$A\left(\mathbf{U}_\pi, \tilde{\mathbf{U}}_\pi\right) = \sum_{ij} \left(\tilde{\mathbf{U}}_\pi \mathbf{H}\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi\mathbf{H}^{\mathrm{T}} + \lambda\mathbf{D}_\pi^u\tilde{\mathbf{U}}_\pi + \mathbf{1}_m\mathbf{1}_m^{\mathrm{T}}\tilde{\mathbf{U}}_\pi\mathbf{\Lambda}_\pi\right)_{ij}\frac{(\mathbf{U}_\pi)_{ij}^2}{(\tilde{\mathbf{U}}_\pi)_{ij}}$$
$$- 2\sum_{ij}\left(\mathbf{X}_\pi\mathbf{V}_\pi\mathbf{H}^{\mathrm{T}} + \lambda\mathbf{W}_\pi^u\tilde{\mathbf{U}}_\pi + \mathbf{1}_m\mathbf{1}_k^{\mathrm{T}}\mathbf{\Lambda}_\pi\right)_{ij}(\tilde{\mathbf{U}}_\pi)_{ij}\left(1 + \log\frac{(\mathbf{U}_\pi)_{ij}}{(\tilde{\mathbf{U}}_\pi)_{ij}}\right)$$

is an auxiliary function for $L(\mathbf{U}_\pi)$. Furthermore, it is a convex function with respect to $\mathbf{U}_\pi$ and has a global minimum.

Theorem 2 can be proved similarly as (Ding et al. 2006) by validating $A(\mathbf{U}_\pi, \tilde{\mathbf{U}}_\pi) \geq L(\mathbf{U}_\pi)$, $A(\mathbf{U}_\pi, \mathbf{U}_\pi) = L(\mathbf{U}_\pi)$, and the Hessian matrix $\nabla\nabla_{\mathbf{U}_\pi}A(\mathbf{U}_\pi, \tilde{\mathbf{U}}_\pi) \succeq 0$. Due to limited space, we omit the details of the validation here. Based on Theorem 2, we can minimize $A(\mathbf{U}_\pi, \tilde{\mathbf{U}}_\pi)$ with respect to $\mathbf{U}_\pi$ with $\tilde{\mathbf{U}}_\pi$ fixed. Set $\nabla_{\mathbf{U}_\pi}A(\mathbf{U}_\pi, \tilde{\mathbf{U}}_\pi) = \mathbf{0}$ we obtain the following updating formula

$$\mathbf{U}_\pi \leftarrow \tilde{\mathbf{U}}_\pi \circ \sqrt{\frac{\left[\mathbf{X}_\pi\mathbf{V}_\pi\mathbf{H}^{\mathrm{T}} + \lambda\mathbf{W}_\pi^u\tilde{\mathbf{U}}_\pi + \mathbf{1}_m\mathbf{1}_k^{\mathrm{T}}\mathbf{\Lambda}_\pi\right]}{\left[\tilde{\mathbf{U}}_\pi\mathbf{H}\mathbf{V}_\pi^{\mathrm{T}}\mathbf{V}_\pi\mathbf{H}^{\mathrm{T}} + \lambda\mathbf{D}_\pi^u\tilde{\mathbf{U}}_\pi + \mathbf{1}_m\mathbf{1}_m^{\mathrm{T}}\tilde{\mathbf{U}}_\pi\mathbf{\Lambda}_\pi\right]}}$$

which is consistent with the updating formula derived from the KKT condition aforementioned. By Lemma 1 and Theorem 2, for each iteration of updating $\mathbf{U}_\pi$, we have $L(\mathbf{U}_\pi^0) = Z(\mathbf{U}_\pi^0, \mathbf{U}_\pi^0) \geq Z(\mathbf{U}_\pi^1, \mathbf{U}_\pi^0) \geq Z(\mathbf{U}_\pi^1, \mathbf{U}_\pi^1) = L(\mathbf{U}_\pi^1) \geq \ldots \geq L(\mathbf{U}_\pi^{maxIter})$. So $L(\mathbf{U}_\pi)$ is monotonically decreasing during iteration. Since the objective function in Equation (7) is lower bounded by 0, the correctness and convergence of Theorem 1 and Algorithm 1 are established.

Experiments

Data Sets and Evaluation Criteria

We evaluate the proposed GCMF approach on nine cross-domain data sets generated from two benchmark text data sets 20-Newsgroups[1] and Reuters-21578[2], which are widely adopted for transfer learning evaluation (Dai et al. 2007; Ling et al. 2008; Pan et al. 2011; Zhuang et al. 2011).

20-Newsgroups has approximately 20,000 documents distributed evenly in 20 different subcategories. The corpus contains four top categories *comp*, *rec*, *sci* and *talk*. Each top category has four subcategories, which are split into two groups *A* and *B*, as listed in Table 2. Following the approach in (Dai et al. 2007), to set up one data set (including source domain and target domain), we randomly select two out of the four top categories, denoted by *P* and *Q*. Then, groups *A* in *P* (denoted by *P.A*) and *Q* (denoted by *Q.A*) are merged as the source domain, while *P.B* and *Q.B* are merged as the target domain. Clearly, for each example in the source domain and target domain, its class label is either *P* or *Q*. In this way, we generate six data sets *comp vs rec* (i.e., *P* = *comp*, *Q*=*rec*; we can explain the other pairs in the same way), *comp vs sci*, *comp vs talk*, *rec vs sci*, *rec vs talk*, and *sci vs talk*. For fair comparison, the six data sets are constructed using a preprocessed version of 20-Newsgroups in (Zhuang et al. 2011).

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

[2]http://www.daviddlewis.com/resources/testcollections/reuters21578/

Table 2: Top categories and their subcategories. Each top category is partitioned into two groups *A* and *B*.

Top Categories	Subcategories
comp	*(A) comp.graphics, comp.os.ms-windows.misc*
	(B) comp.sys.ibm.pc.hardware, comp.sys.mac.hardware
rec	*(A) rec.autos, rec.motorcycles*
	(B) rec.sport.baseball, rec.sport.hokey
sci	*(A) sci.crypt, sci.electronics*
	(B) sci.med, sci.space
talk	*(A) talk.politics.guns, talk.politics.mideast*
	(B) talk.politics.misc, talk.religion.misc

Reuters-21578 is another widely used data set for evaluating learning algorithms. It contains five top categories and many subcategories. For easy comparison, we use a preprocessed version adopted in previous work (Gao et al. 2008; Zhuang et al. 2010), which contains three cross-domain data sets *orgs vs people*, *orgs vs place* and *people vs place*.

We use *Accuracy* as the evaluation criteria, as it is widely adopted in the literature (Pan et al. 2011; Zhuang et al. 2011)

$$Accuracy = \frac{|\{\mathbf{x} : \mathbf{x} \in \mathcal{D}_t \wedge f(\mathbf{x}) = y(\mathbf{x})\}|}{|\{\mathbf{x} : \mathbf{x} \in \mathcal{D}_t\}|}$$

where $y(\mathbf{x})$ is the groundtruth label of example $\mathbf{x}$ while $f(\mathbf{x})$ is the label predicted by the classification algorithm.

Baseline Methods and Parameter Settings

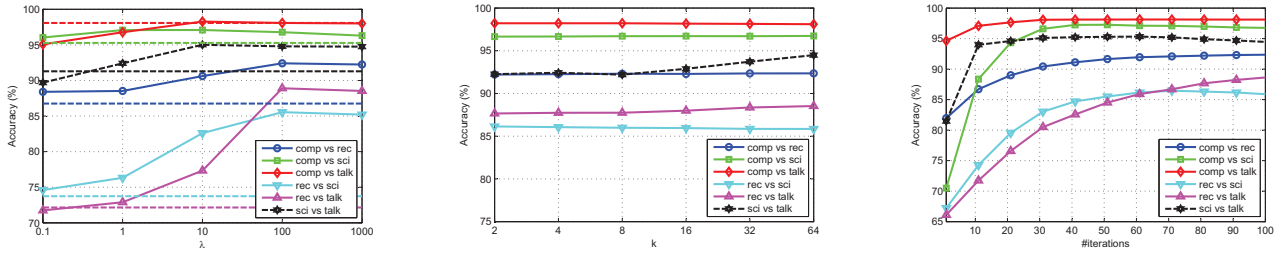
We compare our proposed GCMF approach with six state-of-the-art baselines, as shown below.

- Unsupervised learning method Nonnegative Matrix Factorization (NMF); supervised learning methods Logistic Regression (LG) and Support Vector Machine (SVM); semi-supervised learning method Transductive Support Vector Machine (TSVM) (Joachims 1999).
- State-of-the-art transfer learning methods Matrix Trifactorization based Classification (MTrick) (Zhuang et al. 2010) and Dual Knowledge Transfer (DKT) (Wang et al. 2011). These methods are applied in a transductive configuration, i.e., trained on all available data and tested on target domain data, using their optimal parameter settings.

GCMF involves a few parameters. The number of classes c is 2 in our data sets. We set the iteration number $maxIter$ as 100. There are four more parameters: regularization parameters λ and γ, number of feature clusters k, number of nearest neighbors p. In the coming sections, we provide detailed analysis on parameter sensitivity, where we evaluate $\lambda = \gamma$ with the values among $\{0.1, 1, 10, 100, 1000\}$. Similarly, we evaluate k among $\{2, 4, 8, 16, 32, 64\}$ and p among $\{2, 4, 6, 8, 10\}$. From the analysis, we can see that GCMF achieves stable performance with a wide range of parameter values. When comparing with the baselines, we use the following parameter settings: $\lambda = \gamma = 100, k = 64, p = 10$.

Experimental Results

The best results of the baselines and GCMF on all the data sets are presented in Table 3. Note that Algorithm 1 involves



(a) Classification accuracy with respect to regularization parameters $\lambda = \gamma$.

(b) Classification accuracy with respect to #feature clusters k.

(c) Classification accuracy with respect to #iterations.

Figure 2: Parameter sensitivity and convergence of GCMF on the cross-domain data sets generated from 20-Newsgroups.

Table 3: Average classification accuracy (%) on the cross-domain data sets (10 repeated experiments).

Data Set	NMF	LG	SVM	TSVM	MTrick	DKT	GCMF
comp vs rec	79.11	81.71	81.04	82.23	86.75	85.83	**92.36±0.00**
comp vs sci	65.88	67.89	64.95	69.75	95.27	92.44	**96.75±0.01**
comp vs talk	56.62	93.05	89.42	95.37	98.08	98.04	**98.11±0.00**
rec vs sci	75.10	67.22	67.44	70.86	73.74	70.48	**85.84±0.01**
rec vs talk	58.80	65.95	65.29	67.84	72.15	68.96	**88.61±0.07**
sci vs talk	61.75	78.42	73.76	80.22	90.08	91.28	**94.35±0.39**
orgs vs people	63.29	74.25	74.88	73.80	80.96	81.46	**85.48±0.04**
orgs vs place	72.63	69.99	71.89	69.89	76.41	76.70	**77.82±0.13**
people vs place	60.49	59.05	58.06	58.43	67.87	66.11	**68.69±0.13**
Average	65.96	73.06	71.86	74.27	82.37	81.26	**87.56±0.09**

random initializations. Therefore we run GCMF 10 times and then report its average classification accuracy with standard deviation on the test set (target domain unlabeled data).

From Table 3, we can see that Non-transfer methods cannot perform well on most data sets. NMF performs poorly since the samples in these data sets are not well separated originally. For LG and SVM, the classifiers trained on the source domain data cannot discriminate the target domain data. Transductive method TSVM outperforms NMF, LG and SVM in most cases. However, its performance is far from satisfactory. Generally, traditional non-transfer methods treat data from different domains as if they were drawn from a homogenous distribution, which explains why their performance is not optimal on the cross-domain data sets.

Transfer learning methods MTrick and DKT work better than the non-transfer methods. These transfer learning methods try to build classifier by leveraging the knowledge from the source domain. However, these methods individually have not reached the best performance yet for some of the data sets (e.g., *rec vs sci*, *rec vs talk*). The reason is that they each build models by exploiting only certain nature of the data, that is, either the empirical likelihood or the geometric structure. They will be at risk when the data are governed by the specific nature which they do not explore.

The proposed GCMF approach obtains much better performance than all the baselines on all data sets with statistical significance. The average classification accuracy improvement is 5.2% on all data sets, which means 29.4% error reduction compared to the best baseline. This proves that by maximizing the empirical likelihood and preserving the geometric structure simultaneously, GCMF can successfully discover the common latent factors as a robust bridge for knowledge transfer. This facilitates smooth transfer learning and results in high classification accuracy.

Parameter Sensitivity

To study the parameter sensitivity of GCMF, we tune the parameters one at each time while fixing others to their optimal values. We do the experiments on all the data sets, but just report the results on the six data sets generated from 20-Newsgroups for simplicity and clarity.

Figure 2a shows the average classification accuracy of GCMF under varying values of λ and γ. We can find that GCMF performs very well and steadily when λ and γ span over a wide range, i.e., $[10, 1000]$. The dashed curves with corresponding colors represent the best performance obtained by the baselines on these data sets, which are below the GCMF curves in most cases. This validates that preserving the geometric structure can indeed help discover more transferrable common latent factors which facilitate smooth knowledge transfer. Similarly, we also studied the parameter sensitivity of p (the number of nearest neighbors) and found that GCMF reaches the best performance when $p \in [4, 10]$, which is consistent with (Cai et al. 2009).

Figure 2b shows the average classification accuracy of GCMF under varying values of k, the number of feature clusters. We can see that GCMF performs steadily when k takes value in a wide range, i.e., $[16, 64]$.

Convergence and Time Complexity

We investigate the convergence of GCMF empirically. Figure 2c shows the average classification accuracy with respect to the number of iterations. We can find that the average classification accuracy of GCMF increases steadily with more iterations and then converges after 100 iterations.

Finally, we check the time complexity of GCMF empirically on the six data sets generated from 20-Newsgroups. We find that GCMF runs very efficiently and takes no more than 100 seconds (when $k = 64$) on each data set containing approximately 8,000 documents and 25,800 features.

Conclusion

In this paper, we propose a novel transfer learning approach to maximize the empirical likelihood and preserve the geometric structure simultaneously. This solves the shortcomings of most existing transfer learning methods which focus only one aspect of the data. Our approach can also serve as a general framework to incorporate various prior knowledge so long as it can be represented by a graph structure. We compare our proposed approach against six state-of-the-art baselines over the widely adopted data sets. The experiments prove that our approach significantly outperforms all the baselines under different settings. After validating our approach on the public data sets, we plan to apply it to some large-scale and real-life applications in the future.

Acknowledgments

The work is supported by the National HeGaoJi Key Project (No. 2010ZX01042-002-002-01), the National Basic Research Program of China (No. 2009CB320706), and the National Natural Science Foundation of China (No. 61050010, No. 61073005, No. 60972096).

References

Belkin, M., and Niyogi, P. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Proceedings of Neural Information Processing Systems 15*, NIPS.

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

Cai, D.; He, X.; Wang, X.; Bao, H.; and Han, J. 2009. Locality preserving nonnegative matrix factorization. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, IJCAI.

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

Ding, C.; Li, T.; Peng, W.; and Park, H. 2006. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD.

Ding, C.; Li, T.; and Peng, W. 2006. Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence, chi-square statistic, and a hybrid method. In *Proceedings of the 21st AAAI Conference on Artificial Intelligence*, AAAI.

Gao, J.; Fan, W.; Jiang, J.; and Han, J. 2008. Knowledge transfer via multiple model local structure mapping. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD.

Gu, Q., and Zhou, J. 2009. Co-clustering on manifolds. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD.

Gu, Q.; Ding, C.; and Han, J. 2011. On trivial solution and scale transfer problems in graph regularized nmf. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, IJCAI.

Gupta, S. K.; Phung, D.; Adams, B.; Tran, T.; and Venkatesh, S. 2010. Nonnegative shared subspace learning and its application to social media retrieval. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD.

Joachims, T. 1999. Transductive inference for text classification using support vector machines. In *Proceedings of the 16th International Conference on Machine Learning*, ICML.

Lee, D. D., and Seung, H. S. 2000. Algorithms for non-negative matrix factorization. In *Proceedings of Neural Information Processing Systems 14*, NIPS.

Ling, X.; Dai, W.; Xue, G.-R.; Yang, Q.; and Yu, Y. 2008. Spectral domain-transfer learning. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD.

Long, M.; Cheng, W.; Jin, X.; Wang, J.; and Shen, D. 2010. Transfer learning via cluster correspondence inference. In *Proceedings of the 10th IEEE International Conference on Data Mining*, ICDM.

Long, M.; Wang, J.; Ding, G.; Cheng, W.; Zhang, X.; and Wang, W. 2012. Dual transfer learning. In *Proceedings of the 12th SIAM International Conference on Data Mining*, SDM.

Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22:1345–1359.

Pan, S. J.; Ni, X.; Sun, J.-T.; Yang, Q.; and Chen, Z. 2010a. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on World Wide Web*, WWW.

Pan, W.; Xiang, E.; Liu, N. N.; and Yang, Q. 2010b. Transfer learning in collaborative filtering for sparsity reduction. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, AAAI.

Pan, S. J.; Tsang, I. W.; Kwok, J. T.; and Yang, Q. 2011. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks* 22(2):199–210.

Singh, A. P., and Gordon, G. J. 2008. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD.

Wang, C., and Mahadevan, S. 2009. Manifold alignment without correspondence. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, IJCAI.

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

Wang, H.; Huang, H.; Nie, F.; and Ding, C. 2011. Cross-language web page classification via dual knowledge transfer using nonnegative matrix tri-factorization. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR.

Zhu, X., and Lafferty, J. 2005. Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML.

Zhu, Y.; Chen, Y.; Lu, Z.; Pan, S. J.; Xue, G.-R.; Yu, Y.; and Yang, Q. 2011. Heterogeneous transfer learning for image classification. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, AAAI.

Zhuang, F.; Luo, P.; Xiong, H.; He, Q.; Xiong, Y.; and Shi, Z. 2010. Exploiting associations between word clusters and document classes for cross-domain text categorization. In *Proceedings of the 10th SIAM International Conference on Data Mining*, SDM.

Zhuang, F.; Luo, P.; Shen, Z.; He, Q.; Xiong, Y.; Shi, Z.; and Xiong, H. 2011. Mining distinction and commonality across multiple domains using generative model for text classification. *IEEE Transactions on Knowledge and Data Engineering* 99(PrePrints).