Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence

Domain Adaptation with Topical Correspondence Learning

Zheng Chen and Weixiong Zhang
Department of Computer Science and Engineering
Washington University in St. Louis, St. Louis, MO
{zheng.chen, weixiong.zhang}@wustl.edu

Abstract

A serious and ubiquitous issue in machine learning is the lack of sufficient training data in a domain of interest. Domain adaptation is an effective approach to dealing with this problem by transferring information or models learned from related, albeit distinct, domains to the target domain. We develop a novel domain adaptation method for text document classification under the framework of Non-negative Matrix Factorization. Two key ideas of our method are to construct a latent topic space where a topic is decomposed into common words shared by all domains and words specific to individual domains, and then to establish associations between words in different domains through the common words as a bridge for knowledge transfer. The correspondence between cross-domain topics leads to more coherent distributions of source and target domains in the new representation while preserving the predictive power. Our new method outperformed several state-of-the-art domain adaptation methods on several benchmark datasets.

1 Introduction

The conventional discriminative learning implicitly assumes that training and test data follow the same distribution. This assumption, however, seldom holds in reality. In contrary, it is usually expensive to obtain an adequate amount of training data that follow the same distribution of the test data. On the other hand, there often exists a sufficient amount of labeled data from related, albeit different, domains. The objective of domain adaptation or transfer learning is to exploit such labeled data in a *source* domain to accurately predict the labels of test data in a *target* domain where little or no label information is available.

Several domain adaptation methods have been developed recently for applications in diverse areas, such as text document classification [Dai *et al.*, 2007a], computer visions [Saenko *et al.*, 2010] and computational biology [Liu *et al.*, 2008]. In text document classification, whose primary objective is to predict the classes of given documents, training and test documents may be from different domains. For example, in the *20Newsgroups* data that have been widely used in text document classification, the task *rec vs sci* is to classify recreation and science related articles. Words like "cars" and "engine" as well as "doctor" and "disease" are among the most discriminative words when the documents are drawn from the *automobile* and *medical* domains. However, in the *sports* and *space* domains, which are subcategories of *rec* and *sci*, respectively, the most discriminative words become "game" and "team" as well as "moon" and "earth". If a discriminative model is trained using documents from one domain, the model typically performs poorly in other domains. This is primarily because domain-specific vocabularies are used to describe distinct concepts in different domains (i.e., subcategories here), e.g., concepts related to recreation versus that related to science. More critically, some of the most discriminative words in one domain may be missing entirely in other domains, resulting in incomparable feature spaces between training and test data.

Topical models [Xue *et al.*, 2008; Jin *et al.*, 2011] and Non-negative Matrix Factorization (NMF) based methods [Lee *et al.*, 1999; Zhuang *et al.*, 2011; Long *et al.*, 2012a; 2012b] have been proposed to alleviate the problem of divergent feature spaces by learning a joint representation that integrates features in the source and target domains. In text document classification, for example, it is natural to consider a document as a linear combination of semantic topics. Ideally, the class of a given document can be uniquely determined by a composition of the topics described, presumably semantically independent of the domains considered. If we can project a document from the word space to a semantics space, where each axis corresponds to a particular topic, documents from different domains can be unified in a common representation. A major challenge here is to find a projection that is not only coherent in multiple domains but also able to preserve the discriminative power after the projection.

In this paper, we proposed Topical Correspondence Learning (TCL), a novel domain adaptation method. The central element of TCL is an optimization problem, formulated as a joint Non-negative Matrix Tri-Factorization (NMTF), with the objective of simultaneously learning a topical representation and conditional probabilities of document classes. The key idea of TCL stems from the observation that a semantic topic is typically comprised of two types of words, i.e., common and domain-specific words. Even though many domain-specific words appear only in one domain, which seems to

be useless for cross-domain classification, they are invaluable for discriminating classes within the domain. Common words can then serve to bridge the gap between domain-specific words for transferring knowledge across domains. TCL models the domain dependency of words in each topic by decomposing words into common and domain-specific parts in such a way that domain-specific words are explicitly associated with common words in the same topic. Comparing with the existing NMF based methods, TCL promotes the correspondence between topics in different domains and thus draws the distributions of the source and target domains closer while preserving the discriminative power in the new topical representation. TCL incorporates the label information into the optimization framework in a semi-supervised manner by directly modeling the posterior probability of labels so that no additional classifier is needed to make predictions. The TCL optimization problem is solved with an efficient iterative updating algorithm with theoretical convergence guarantees. We empirically evaluated the performance of TCL on three well-adopted benchmark datasets. Our approach outperforms several state-of-the-art methods for cross-domain document classification. Our results show that the learned topics by TCL are semantically meaningful and are useful for understanding the commonalty and difference among different domains.

In the following, we first discuss the basic notations and problem settings in Section 2. We then describe the new method along with a theoretical analysis in Section 3. Experiment results are presented in Section 4, and related works are reviewed in Section 5. We conclude in Section 6.

2 Notation and Setting

We use bold capital letters, e.g., $\mathbf{X}$ and $\mathbf{Y}$, for matrices and bold lowercase letters, e.g., $\mathbf{x}$ and $\mathbf{y}$, for vectors. We denote $(\mathbf{X})_{i:}$ the i-th row of matrix $\mathbf{X}$, $(\mathbf{X})_{:j}$ the j-th column of $\mathbf{X}$, and $(\mathbf{X})_{ij}$ the entry at the i-th row and j-th column of $\mathbf{X}$.

Let $\{\mathcal{D}_d \mid d \in [1, D]\}$ be a family of D domains. Following the paradigm of transductive transfer learning, we assume that documents are from two types of domains, source $\mathcal{S}$ where documents are labeled, and target $\mathcal{T}$ where documents are unlabeled. For each domain $\mathcal{D}_d$, we sample an $m \times n_d$ matrix of data $\mathbf{X}_d = \{\mathbf{x}_1, ..., \mathbf{x}_{n_d}\}$, where $\mathbf{x}_i^{(d)}$ is a feature vector in the shared feature space with m words. We assume that the raw feature representation only involves positive numbers, e.g. bag-of-words or term frequency-inverse document frequency (tf-idf) [Salton and Buckley, 1988]. Suppose that there are c classes. Each source domain $\mathcal{D}_d \in \mathcal{S}$ is associated with a $n_d \times c$ label matrix $\mathbf{Y}_d = \{\mathbf{y}_1, ..., \mathbf{y}_{n_d}\}$, where if document $\mathbf{x}_i^{(d)}$ is in class $j \in [1, c]$ then $\mathbf{y}_{ij}^{(d)} = 1$, or $\mathbf{y}_{ij}^{(d)} = 0$, otherwise. Our goal is to learn a classifier $h \in \mathcal{H}$ given the labeled set $\{(\mathbf{X}_d, \mathbf{Y}_d) \mid \mathcal{D}_d \in \mathcal{S}\}$ to accurately predict the labels of an unlabeled sets $\{\mathbf{X}_d \mid \mathcal{D}_d \in \mathcal{T}\}$.

3 The TCL Method

We begin with a brief review of Non-negative Matrix Tri-Factorization (NMTF) [Ding *et al.*, 2006] and its extension to domain adaptation [Li *et al.*, 2010]. We then present our TCL method, which preserves the modeling power of NMTF and exploits the correspondence of topics to support cross-domain knowledge transfer.

3.1 Non-negative Matrix Tri-Factorization

Given a non-negative $m \times n$ word-document matrix $\mathbf{X}$, NMTF is to approximate $\mathbf{X}$ by the product of three non-negative matrices: $\mathbf{X} \approx \mathbf{W}\mathbf{H}\mathbf{V}^T$, where $\mathbf{W}$ is an $m \times k$ word-topic matrix, each column of $\mathbf{W}$ defines a latent topic, and $(\mathbf{W})_{ij}$ describes the contribution of the i-th word to the j-th topic; $\mathbf{V}$ is an $n \times c$ document-class matrix representing soft membership of documents in c classes; and $\mathbf{H}$ is a $k \times c$ matrix indicating the association between the document classes and latent topics. The non-negativity constraint on all factors stems from the intuition that the contribution of topics should be additive rather than subtractive. This property leads to interpretability and sparse solutions to the problem [Lee *et al.*, 1999]. Specifically, NMTF solves the following optimization problem through a set of iterative update,

$$\begin{aligned} \min_{\mathbf{W},\mathbf{H},\mathbf{V}\geq 0} & \frac{1}{2}\left\|\mathbf{X} - \mathbf{W}\mathbf{H}\mathbf{V}^T\right\|_F^2 \\ s.t. \quad & \mathbf{W}^T\mathbf{1}_m = \mathbf{1}_k, \mathbf{V}\mathbf{1}_c = \mathbf{1}_n, \end{aligned} \tag{1}$$

where $\|.\|_F$ is the Frobenius norm and $\mathbf{1}_u$ is a $u \times 1$ column vector with all entries equal to 1. The equality constraints on $\mathbf{W}$ and $\mathbf{V}$ avoid ambiguity of the factorization and, furthermore, provide $\mathbf{W}$ and $\mathbf{V}$ with a probability interpretation, i.e., $(\mathbf{W})_{ij}$ becomes the posterior probability of the i-th word belonging to the j-th topic and $(\mathbf{V})_{ij}$ becomes the posterior probability of the i-th document being in class j.

In order to apply NMTF to a corpus of documents $\{\mathbf{X}_d\}$ from multiple domains $\{\mathcal{D}_d\}$, the minimization problem in (1) can be extended to,

$$\min_{\mathbf{W}_d,\mathbf{H},\mathbf{V}_d\geq 0} \frac{1}{2}\sum_d^D \left\|\mathbf{X}_d - \mathbf{W}_d\mathbf{H}\mathbf{V}_d^T\right\|_F^2, \tag{2}$$

where $\mathbf{W}_d$ is the word-topic matrix, or the topical model, and $\mathbf{V}_d$ the document-class matrix in the d-th domain. From the perspective of classification, $\mathbf{W}_d$ can be viewed as the coordinates of a k-dimensional projection space of topics, and $\mathbf{H}\mathbf{V}_d^T$ together can be considered as a new representation of the documents in the projected space defined by $\mathbf{W}_d$. Here the common factor, $\mathbf{H}$, accommodates the notion that the association between topics and document classes is domain independent. $\mathbf{H}$ serves as a topic-level bridge among domains by implicitly associating topics from different domains using the same coefficients. As a result, the label of a given document, $\mathbf{x}_i^{(d)}$, in domain d can be determined by:

$$f(\mathbf{x}_i^{(d)}) = \operatorname*{argmax}_{j\in[1,c]} (\mathbf{V}_d)_{ij}.$$

3.2 Topical Correspondence Learning

The primary objective of an NMF-based method for transductive learning is to learn a topic model, $\mathbf{W}_d$, to define the coordinates of a new representation in which the distributions of source and target documents are drawn closer than

before the projection. Even though the topics, $(\mathbf{W}_{d_1})_{:i}$ and $(\mathbf{W}_{d_2})_{:i}$, with the same column index i across two domains, $\mathcal{D}_{d_1}$ and $\mathcal{D}_{d_2}$, are implicitly associated through $(\mathbf{H})_{i:}$, there is no guarantee that the coordinates of the new representation are closely related because a set of topics are learned independently in each domain . A possible remedy is to use a common word-topic matrix $\tilde{\mathbf{W}}$ across all domains such that $\mathbf{X}_d \approx \tilde{\mathbf{W}}\mathbf{H}\mathbf{V}_d^T$. Although it guarantees that all domains share the same topic coordinates, this will seriously undermine the modeling power of NMF because it disregards the difference between vocabularies in different domains and thus results in an incoherent new representation. To overcome this problem, some recent works [Long *et al.*, 2012a; Gupta *et al.*, 2011] partition the word-topic matrix $\tilde{\mathbf{W}}$ into two parts $[\mathbf{U}, \mathbf{W}_d]$, where $\mathbf{U}$ is an $m \times k_1$ matrix representing the k_1 common topics shared by all domains and $\mathbf{W}_d$ is an $m \times k_2$ matrix representing the k_2 domain-specific topics unique to the d-th domain. The objective of the extended NMTF [Long *et al.*, 2012a] can be formulated as,

$$\min_{\mathbf{U},\mathbf{W}_d,\mathbf{H},\mathbf{V}_d \geq 0} \frac{1}{2}\sum_{d}^{D} \left\|\mathbf{X}_d - [\mathbf{U}, \mathbf{W}_d]\mathbf{H}\mathbf{V}_d^T\right\|_F^2. \quad (3)$$

Among all $k = k_1 + k_2$ topics in each domain, at least k_1 coordinates are exactly aligned. The remaining k_2 topics are used to properly model the domain-specific words to avoid negative transfer [Gupta *et al.*, 2011].

A key observation here is that a domain-specific word is unlikely to be included in a common topic in $\mathbf{U}$, because if this happened, any document associated with the topic in the other domains would have to include that domain-specific word, thus increasing the reconstruction error as a result. In other words, topics defined in $\mathbf{W}_d$ tend to contain all domain-specific words. In fact, even though the distributions of different domains are closer in the subspace defined by $\mathbf{U}$, the distributions become more distant under $\mathbf{W}_d$ because the corresponding coordinates are almost always orthogonal to one another. Due to the lack of association between cross-domain topics in $\mathbf{W}_d$, domain-specific words will not be particularly useful for a cross-domain discriminative learner. As discussed in Section 1, domain-specific words, however, still contain valuable knowledge because many of them may be among the most discriminative features.

Domain-specific words can be exploited. The key idea, which forms the cornerstone of this research is to build a correspondence between cross-domain topics in $\mathbf{W}_d$ by associating domain-specific words with common words that are shared by all domains. From a document generative perspective, we assume that when we choose a word from a certain topic to generate a document, either a common word or a domain-specific word is selected. For each topic, the objective is to learn a decomposition of a common vocabulary $(\mathbf{U})_{:i}$ and d domain-specific vocabularies $(\mathbf{W}_d)_{:i}$. Specifically, we construct a one-to-one correspondence between cross-domain topics, $\mathbf{W}_d$, and the shared topics $\mathbf{H}$, arriving

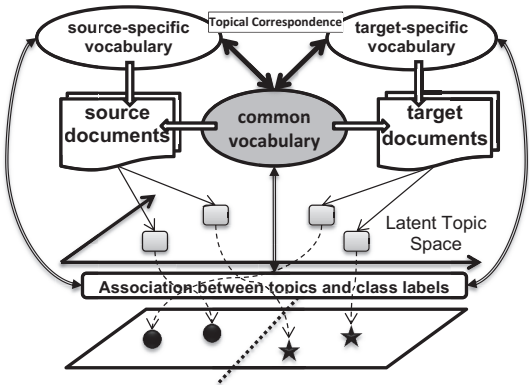

Figure 1: Topical Correspondence between common and domain-specific words.

at minimizing the following objective function,

$$\mathcal{O} = \frac{1}{2}\sum_{d}^{D} \left\|\mathbf{X}_d - [\alpha\mathbf{U}, (1-\alpha)\mathbf{W}_d]\begin{bmatrix}\mathbf{H}\\ \mathbf{H}\end{bmatrix}\mathbf{V}_d^T\right\|_F^2 \quad (4)$$
$$s.t. \quad \mathbf{U}^T\mathbf{1}_m = \mathbf{1}_k, \mathbf{W}_d^T\mathbf{1}_m = \mathbf{1}_k, \mathbf{V}_d\mathbf{1}_c = \mathbf{1}_n.$$

Here the factor of association between topics and document classes is a vertical duplication of $k \times c$ matrix $\mathbf{H}$. $\mathbf{U}$ and $\mathbf{W}_d$, which are word-topic matrices of the same size, $m \times k$, are comprised of common and domain-specific words respectively. The hyperparameter $\alpha \in [0, 1]$ determines the probability of choosing a common or domain-specific word. By enforcing the same $\mathbf{H}$ on $\mathbf{U}$ and $\mathbf{W}_d$, we in fact force the reconstruction of documents in a way that, whenever a common topic $(\mathbf{U})_{i:}$ is chosen, a proportional amount, $\frac{1-\alpha}{\alpha}$, of $(\mathbf{W}_d)_{:i}$ has also been chosen in the d-th domain as well. Accordingly we explicitly associate topics, $(\mathbf{W}_d)_{:i}$, of different domains through the common topic, $(\mathbf{U})_{:i}$, by the same coefficients $(\mathbf{H})_{i:}$. Due to this feature, we name our method Topical Correspondence Learning. Figure 1 sketches the intuition behind the proposed method.

3.3 The Algorithm

The TCL optimization problem in Eq. (4) is not jointly convex in all variables. Therefore, it is unrealistic to find global minima. Therefore we derive an iterative algorithm to search for local optima by alternatively updating one variable and fixing the others in one step of the search. Denote $\mathbf{P}_d = (\alpha\mathbf{U} + (1-\alpha)\mathbf{W}_d)$. The optimization problem in Eq. (4) can be solved using the following update rules,

$$(\mathbf{U})_{ij} \leftarrow (\mathbf{U})_{ij}\frac{(\sum_d^D \mathbf{X}_d\mathbf{V}_d\mathbf{H}^T)_{ij}}{(\sum_d^D \mathbf{P}_d\mathbf{H}\mathbf{V}_d^T\mathbf{V}_d\mathbf{H}^T)_{ij}} \quad (5)$$

$$(\mathbf{W}_d)_{ij} \leftarrow (\mathbf{W}_d)_{ij}\frac{(\mathbf{X}_d\mathbf{V}_d\mathbf{H}^T)_{ij}}{(\mathbf{P}_d\mathbf{H}\mathbf{V}_d^T\mathbf{V}_d\mathbf{H}^T)_{ij}} \quad (6)$$

$$(\mathbf{V}_d)_{ij} \leftarrow (\mathbf{V}_d)_{ij}\frac{(\mathbf{X}_d^T\mathbf{P}_d\mathbf{H})_{ij}}{(\mathbf{V}_d\mathbf{H}^T\mathbf{P}_d^T\mathbf{P}_d\mathbf{H})_{ij}} \quad (7)$$

$$(\mathbf{H})_{ij} \leftarrow (\mathbf{H})_{ij}\frac{(\sum_d^D \mathbf{P}_d^T\mathbf{X}_d\mathbf{V}_d)_{ij}}{(\sum_d^D \mathbf{P}_d^T\mathbf{P}_d\mathbf{H}\mathbf{V}_d^T\mathbf{V}_d)_{ij}} \quad (8)$$

Then we normalize the columns of $\mathbf{U}$ and $\mathbf{W}_d$, and the rows of $\mathbf{V}_d$ to l_1-norm to satisfy the equality constrains.

A proper initialization may lead to a fast convergence. Similar to [Long *et al.*, 2012a; 2012b], we initialize the labels of documents in source domains as true labels, i.e. $\{\mathbf{V}_d = \mathbf{Y}_d\}$ if $\mathcal{D}_d \in \mathcal{S}$. Following the multiplicative updates in Eq. (5) to (8), if $(\mathbf{V}_d)_{ij}$ is initialized to 0, it will remain to be 0. Thus the values of $\mathbf{V}_d$ in the source domain are automatically kept unchanged throughout the execution of the algorithm. We also initialize the labels of data in the target domain by Logistic Regression trained on the data in the source domain.

3.4 Theoretical Analysis

The correctness of the algorithm directly follows the derivation of the updating rules as stated in the following theorem:

Theorem 1. *If the iteration of update rules in Eqs.* (5) *to* (8) *converge, it converges to a local optimal solution.*

Proof. Let $\mathbf{\Lambda}^U$, $\mathbf{\Lambda}_d^W$ and $\mathbf{\Lambda}_d^V$ be the Lagrange multiplier matrices of the equality constraints on $\mathbf{U}$, $\mathbf{W}_d$, and $\mathbf{V}_d$, respectively. Using the KKT conditions from the theory of constrained optimization [Boyd and Vandenberghe, 2004], for any stationary point we have the following equations,

$$\left(\begin{array}{l} \sum_d^D(-\mathbf{X}_d\mathbf{V}_d\mathbf{H}^T + \mathbf{P}_d\mathbf{H}\mathbf{V}_d^T\mathbf{V}_d\mathbf{H}^T) \\ +\mathbf{1}_m\mathbf{1}_m^T\mathbf{U}\mathbf{\Lambda}^U - \mathbf{1}_m\mathbf{1}_k^T\mathbf{\Lambda}^U \end{array} \right)_{ij} (\mathbf{U})_{ij} = 0$$

$$\left(\begin{array}{l} -\mathbf{X}_d\mathbf{V}_d\mathbf{H}^T + \mathbf{P}_d\mathbf{H}\mathbf{V}_d^T\mathbf{V}_d\mathbf{H}^T \\ +\mathbf{1}_m\mathbf{1}_m^T\mathbf{W}_d\mathbf{\Lambda}_d^W - \mathbf{1}_m\mathbf{1}_k^T\mathbf{\Lambda}_d^W \end{array} \right)_{ij} (\mathbf{W}_d)_{ij} = 0$$

$$\left(\begin{array}{l} -\mathbf{X}_d^T\mathbf{P}_d\mathbf{H} + \mathbf{V}_d\mathbf{H}^T\mathbf{P}_d^T\mathbf{P}_d\mathbf{H} \\ +\mathbf{\Lambda}_d^V\mathbf{V}_d\mathbf{1}_c\mathbf{1}_c^T - \mathbf{\Lambda}_d^V\mathbf{1}_n\mathbf{1}_c^T \end{array} \right)_{ij} (\mathbf{V}_d)_{ij} = 0$$

$$\sum_d^D(-\mathbf{P}_d^T\mathbf{X}\mathbf{V}_d + \mathbf{P}_d^T\mathbf{P}_d\mathbf{H}^T\mathbf{V}_d^T\mathbf{V}_d)_{ij}(\mathbf{H})_{ij} = 0$$

Consider the updating rule for $\mathbf{W}_d$ first. If the iteration on Eq. (6) converges, the converged solution $\mathbf{W}_d^*$ satisfies

$$(\mathbf{W}_d^*)_{ij} = (\mathbf{W}_d^*)_{ij}\frac{(\mathbf{X}_d\mathbf{V}_d\mathbf{H}^T + \mathbf{1}_m\mathbf{1}_k^T\mathbf{\Lambda}_d^W)_{ij}}{(\mathbf{P}_d\mathbf{H}\mathbf{V}_d^T\mathbf{V}_d\mathbf{H}^T + \mathbf{1}_m\mathbf{1}_m^T\mathbf{W}_d^*\mathbf{\Lambda}_d^W)_{ij}}$$

Although it is difficult to explicitly compute the Lagrange multiplier $\mathbf{\Lambda}_d^W$, after normalization of $\mathbf{W}_d$, the two terms $\mathbf{1}_m\mathbf{1}_k^T\mathbf{\Lambda}_d^W$ and $\mathbf{1}_m\mathbf{1}_m^T\mathbf{W}_d\mathbf{\Lambda}_d^W$ are in fact equal. They can be safely ignored when considering the convergence. A similar technique can be applied to prove the convergence of $\mathbf{U}$ and $\mathbf{V}_d$. Therefore, this gives rise to the updating rules that satisfies the KKT conditions. □

We have the following theorem regarding the convergence of the iterative updating rules.

Theorem 2. *The objective function $\mathcal{O}$ in* (4) *is non-increasing under the updating rules in Eqs.* (5) *to* (8)*. $\mathcal{O}$ is invariant under these updates if and only if $\mathbf{U}$, $\mathbf{W}_d$, $\mathbf{H}$ and $\mathbf{V}_d$ become stationary.*

Theorem 2 grantees the iterative algorithm to converge to local optima. Due to space limitation, we omit the proof but give a sketch here. To prove Theorem 2, we use the auxiliary function approach used in Expectation-Maximization [Dempster *et al.*, 1977] and NMF [Lee *et al.*, 1999]. The basic idea is to find an auxiliary function $\mathcal{C}(x, x')$ that is an upper bound of the objective function $\mathcal{O}(x)$ and the equality of these two functions holds at $x = x'$. Then we can show that each update in the form of $x^{(t+1)} \leftarrow x^{(t)}$ from Eqs. (5) to (8) is equivalent to the minimization of $x^{(t+1)} = \text{argmin}_x \mathcal{C}(x, x^{(t)})$, which ensures that $\mathcal{O}$ is non-increasing.

Tasks	Source	Target
real vs. sim	{real, sim}-auto	{real, sim}-aviation
auto vs. aviation	sim-{auto, aviation}	real-{auto, aviation}
comp vs. rec	comp.{graphics, os} rec.{autos, motorcycles}	comp.sys.{ibm, mac} rec.sport{baseball, hockey}
comp vs. sci	comp.{graphics, os} sci.{crypt, electronics}	comp.sys.{ibm, mac} sci.{med, space}
comp vs. talk	comp.{graphics, os} politics.{guns, middleeast}	comp.sys.{ibm, mac} talk.{politics.misc, religion}
rec vs. sci	rec.{autos, motorcycles} sci.{crypt, electronics}	rec.sport{baseball, hockey} sci.{med, space}
rec vs. talk	rec.{autos, motorcycles} politics.{guns, middleeast}	rec.sport{baseball, hockey} talk.{politics.misc, religion}
sci vs. talk	sci.{crypt, electronics} politics.{guns, middleeast}	sci.{med, space} talk.{politics.misc, religion}
orgs vs. people	orgs.{...}, people.{...}	orgs.{...}, people.{...}
orgs vs. place	orgs.{...}, place.{...}	orgs.{...}, place.{...}
people vs. place	people.{...}, place.{...}	people.{...}, place.{...}

Table 1: The description of tasks for cross-domain classification.

4 Results

4.1 Datasets Tested

We evaluated the TCL algorithm together with the existing domain adaption algorithms on three benchmark datasets: SRAA[1], 20-Newsgroups[2], and Reuters-21578[3]. We restricted our evaluation to binary classification for clarity and convenience, while TCL can be readily extended to multi-class classification tasks.

SRAA dataset contains $73,218$ UserNet articles from four discussion groups: simulated auto racing (*sim-auto*), simulated aviation (*sim-aviation*), real autos (*real-auto*) and real aviation (*real-aviation*). To create the binary classification task of *sim vs real*, we constructed the source domain of *auto* to contain the *sim-auto* and *real-auto* articles and the target domain of *aviation* to contain the *sim-aviation* and *real-aviation* articles. Likewise, we constructed the task *auto vs aviation* by splitting *sim* and *real* as source and target domains respectively. In each group we sampled around $3,000$ documents, each of which contained at least 20 words, to balance the number of documents across domains. We removed stop words and words appeared in less than 10 documents, resulting in a cohort of $13,472$ documents with $5,982$ features.

20-Newsgroups is a large collection of documents evenly distributed in 20 distinct newsgroups. Each document belongs to one of four top categories and one of the many subcategories. We created one binary classification task for every pair of top

[1]http:// people.cs.umass.edu/ mccallum/data.html

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

[3]http: //www.daviddlewis.com/resources/testcollections/

Data Sets	SVM	LR	mSDA	NMTF	DTL	GCMF	TCL
auto vs. aviation	77.54	80.46	85.35	94.50 ± 0.82	95.34 ± 0.51	94.64 ± 0.02	$\mathbf{97.03 \pm 0.30}$
real vs. sim	75.59	75.42	79.97	93.54 ± 0.53	93.85 ± 0.35	94.55 ± 0.12	$\mathbf{95.12 \pm 0.14}$
comp vs. rec	85.27	87.24	87.99	98.28 ± 0.01	98.25 ± 0.03	$\mathbf{98.28 \pm 0.00}$	98.27 ± 0.02
comp vs. sci	70.81	71.12	79.50	87.00 ± 0.55	89.63 ± 0.44	95.30 ± 0.01	$\mathbf{97.34 \pm 0.21}$
comp vs. talk	95.63	95.93	96.32	86.42 ± 0.49	91.92 ± 0.33	92.60 ± 0.00	$\mathbf{97.18 \pm 0.65}$
rec vs. sci	68.38	68.48	74.09	91.79 ± 0.37	94.77 ± 0.34	95.76 ± 0.03	$\mathbf{97.91 \pm 0.44}$
rec vs. talk	71.05	72.97	69.55	79.39 ± 0.92	89.24 ± 0.95	96.42 ± 0.10	$\mathbf{97.91 \pm 0.10}$
sci vs. talk	75.99	77.94	79.13	73.29 ± 7.82	85.72 ± 5.11	92.29 ± 0.21	$\mathbf{94.63 \pm 0.28}$
org vs. people	71.85	73.18	76.82	66.58 ± 0.21	76.49 ± 0.82	80.50 ± 0.00	$\mathbf{84.32 \pm 0.09}$
org vs places	70.28	70.47	71.52	72.87 ± 0.11	72.79 ± 0.18	74.31 ± 0.02	$\mathbf{78.24 \pm 0.42}$
people vs places	57.10	63.05	61.47	67.80 ± 1.01	68.44 ± 0.90	64.53 ± 0.00	$\mathbf{70.47 \pm 0.31}$
Average	74.50	76.02	78.34	82.86	86.95	89.05	**91.67**

Table 2: Average classification accuracy (%) on various cross-domain classification tasks (All NMF based methods are reported with standard deviation of accuracy on 10 repeated experiments)

categories. From each top category, we selected two subcategories to form the source domain and two others as the target domain so that the classification task in source and target domains were related but their distributions were different. We followed the same procedure for selecting subcategories as in [Long *et al.*, 2012b]. We removed stop words and words occurred in less than 15 documents, resulting in a cohort of $18,774$ documents with $13,781$ features.

Reuters-21578 has been one of the most used test categorization collection. It contains five top categories and many subcategories. Similar to 20-Newsgroups, we used the top categories to construct binary classification tasks, and grouped subcategories into source and target domains. We used the preprocessed data[4] as described in [Pan *et al.*, 2011].

Table 1 lists all the cross-domain tasks that were considered in our evaluation.

4.2 Performance Evaluation and Comparison

To create a baseline for our comparison, we trained a linear Support Vector Machines (SVM) and a Logistic Regression (LR)[5] method on the labeled data in the source domain and tested these two models on the target domain data. We also built a semi-supervised NMTF model using data from both source and target domains. In addition to these three baseline models, we included in our comparison several state-of-the-art domain adaptation methods, including an unsupervised marginalized Stacked Denoising Autoencoder (mSDA) [Chen *et al.*, 2012] and two NMTF based methods, Graph co-regularized Collective Matrix Factorization (GCMF) [Long *et al.*, 2012b] and Dual Transfer Learning (DTL) [Long *et al.*, 2012a].

For SVM and LR, we cross-validated the C parameter on the training data. For NMTF, we implemented the formulation in Eq. (2). The labels in the source domain were included by initializing $\mathbf{V}_d$ as $\mathbf{Y}_d$ and we set the number of topics, k, to 20 because it has a similar effect of the k parameter as in the other NMF-based methods. In mSDA, we stacked 5 layers of features and cross-validated the noise ratio on the training data. We then applied LR to the new representation to evaluate mSDA. For GCMF and DTL, we followed the same parameter selection criteria as in the original papers. Finally, for TCL, we set the number of topics, k, to 10 and the maximal number of iterations for updating to 100 for all tasks, and cross-validated α on training data. For all NMF-based methods, we repeated the experiment 10 times and report the mean and standard deviation of classification accuracy.

[4]http://www.cse.ust.hk/TL/dataset/Reuters.zip

[5]http://www.csie.ntu.edu.tw/ cjlin/liblinear/

4.3 Experiment Results

Performance Evaluation

Table 2 lists the classification accuracies of the methods compared across all domain adaption tasks considered. The new method TCL clearly outperformed the other competing methods in 10 of the 11 tasks. SVM and LR performed poorly on most of these tasks because they were unable to handle the distribution discrepancy between source and target data. As an unsupervised method, mSDA outperformed LR and SVM in most of the cases, indicating that the newly learned feature space can indeed help discriminative learners for cross-domain classification. The semi-supervised NMTF worked surprisingly well and even outperformed mSDA in many cases. Two factors may contribute to the good performance of NMTF. Firstly, the incorporation of labeled data makes the learned latent topics to preserve more discriminative power than unsupervised methods do. Secondly, the classification tasks in our experiment are closely related to the semantic topics, which align well with the assumption for NMTF based methods. As direct extensions to NMTF, both GCMF and DTL achieved better performance than NMTF overall and GCMF performed better than DTL. Finally, except on the task *comp vs rec*, where TCL had a comparable performance with other NMTF based methods, TCL was significantly better than the other methods. On *comp vs rec* task, it is likely the case that domain-dependent topics provided enough discriminative power, while additional knowledge from other domains may not be so useful.

Our implementation of TCL in MATLAB™ is very efficient. It took only ~ 30 seconds on the largest task on a 2.30 GHz quad-core AMD Opteron™ 2376 processor.

Bridge for Knowledge Transfer - Common words

TCL outputs not only the posterior probability of class labels with $\mathbf{V}_d$, but also the posterior probabilities of common and domain-specific words in a given topic, included in $\mathbf{U}$ and $\mathbf{W}_d$, respectively. By sorting the columns of $\mathbf{U}$ and $\mathbf{W}_d$, we can identify the most representative common and domain-specific words in a topic. Table 3 shows an example, listing the representative words from two selected groups

Topic 1	common	*buffalo, east, top, luck, hand*
	source-specific	*car, bike, bmw, engine, ride*
	target-specific	*game, team, season, players, play*
Topic 2	common	*research, system, test, engineering, commercial*
	source-specific	*key, encryption, nsa, keys, security*
	target-specific	*space, nasa, shuttle, moon, orbit*

Table 3: Representative common and domain-specific words in selected topics output by TCL in task *rec vs. sci*

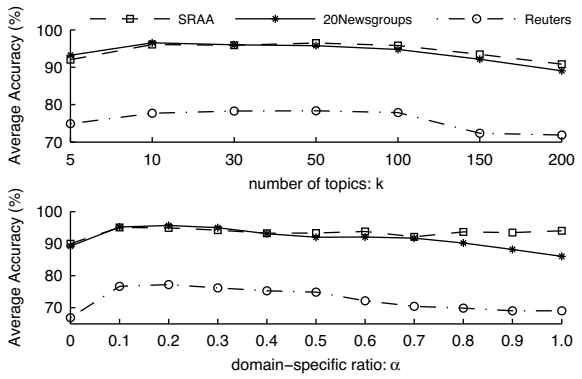

Figure 2: Average accuracy of TCL with varying k and α

of topics in the *rec vs sci* task. Here group 1 contains topics with the same column index of $\mathbf{U}$ and $\mathbf{W}_d$ that are related to the class label *rec*, and topics in group 2 are the ones related to *sci*. We can observe the potential associations of source- and target-specific words through the bridge for knowledge transfer: common words. For instance, in topic 1, it is likely that the words of *car* and *team* are associated with the common word *top*, and in topic 2 the words *space* and *security* are associated with such words as *research* and *system*. This information of correspondence can be useful for knowledge discovery by improving the understanding of the commonalty and difference of different domains.

Effect of Parameters

TCL has two key parameters: the number of latent topics, k, and the weight of domain-specific vocabularies, α. These two parameters are considered to be dataset dependent. In general, k should be large enough to capture the topics in the corpus. However, it should not be too large as it can easily overfit and compromise the ability to generalize to new topics. In the top panel of Figure 2, we vary the values of k from 10 to 200 and plot the average accuracy of TCL on the three datasets separately. We find that the performance of TCL is relatively stable on k in a wide range: $k \in [10, 100]$. Similarly, we plot the average accuracy of TCL on varying values of $\alpha \in [0, 1]$ in the bottom panel of Figure 2. In most cases, $\alpha = 0.1$ achieves a good performance.

5 Related Works and Discussions

The existing domain adaption or transfer learning methods under the transductive settings can be classified into two categories. The algorithms in the first category attempt to draw the distributions of source and target domains closer by assigning different weights to the training instances, i.e., instance re-weighting [Dai *et al.*, 2007b; Mansour *et al.*, 2009]. It has been shown that their performance can asymptotically approach the performance of the optimal in-domain classifier if their underlying assumptions are met.

The algorithms in the second category aim to learn a new shared feature representation in which the source and target distributions are closer than in the original representation. Among these algorithms, Structural Correspondence Learning (SCL) [Blitzer *et al.*, 2006] and deep learning methods [Glorot *et al.*, 2011; Chen *et al.*, 2012] augment the original feature space with newly constructed features. The key idea is to establish correspondences between features cross domains and then add the correspondence information to the new feature representation. Other methods, such as Maximum Mean Discrepancy Embedding [Pan *et al.*, 2008] and Transfer Component Analysis [Pan *et al.*, 2011], project the original features onto a higher level of feature space by assuming the existence of a low dimensional space that is coherent across domains. A major disadvantage of these methods is that they are usually unsupervised; the new representation are not jointly optimized with the label information, which has no guarantee to extract predictive features.

Non-negative Matrix Tri-Factorization (NMTF) [Ding *et al.*, 2006] has recently drawn an extensive attention in document analysis. The NMTF-based algorithms are particularly useful for domain adaptation due to their ability to simultaneously learn a shared latent feature space and incorporate label information to strengthen the discriminative power of new features. In particular, one of the inspiring works of Long *et al.* [2012a], i.e., Dual Transfer Learning (DTL), utilizes an NMTF model as in Eq. (3) so that the marginal and conditional distributions of the source and target domains can be drawn closer at the same time. Comparing to DTL, our new TCL method promotes the coherence of the marginal distribution by explicitly associating commonly shared and domain-specific topics. Closely related to manifold alignment [Wang and Mahadevan, 2011], Graph co-regularized Collective Matrix tri-Factorization (GCMF) [Long *et al.*, 2012b] builds correspondence between cross-domain topics by regularizing the latent feature space to preserve the geometric structure of both words and documents in the original space. It is well suited for tasks where the labels are consistent with the nearest-neighbor graph. However, it still lacks correspondence information across multiple domains without an adequate affinity graph to capture correlations between cross-domain words. The new method TCL, on the other hand, embeds these correlations in the corresponding topics, which essentially drives the distributions of the source and target domains closer than GCMF does in the latent space.

6 Concluding Remarks

We developed a novel domain adaption method under the NMTF framework to simultaneously learn a latent topic representation in the semantic space and the posterior probability of the document labels. We explicitly associated topics across distinct domains by enforcing a one-to-one correspondence between the corresponding columns of the word-topic matrix. Conceptually, we split each topic into several parts:

one for common words and the others for domain-specific words. A key feature of TCL is its ability of utilizing common factors as a bridge to transfer knowledge cross domains. We empirically compared TCL with several baseline models and state-of-the-art methods on three benchmark datasets to demonstrate the efficacy of TCL for cross-domain document classification. As a future direction, we are particularly interested in extending TCL to many application areas, e.g. biology, where the discovery of factors that link and differentiate multiple domains is a primary concern.

Acknowledgment

This work was supported by the National Institutes of Health (R01GM100364) and the National Science Foundation (DBI-0743797).

References

[Blitzer *et al.*, 2006] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of EMNLP06*, pages 120–128, 2006.

[Boyd and Vandenberghe, 2004] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[Chen *et al.*, 2012] M. Chen, Z. Xu, K. Weinberger, and F. Sha. Marginalized denoising autoencoders for domain adaptation. In *Proceedings of the 29th International Conference on Machine learning (ICML)*, 2012.

[Dai *et al.*, 2007a] W. Dai, G.R. Xue, Q. Yang, and Y. Yu. Co-clustering based classification for out-of-domain documents. In *Proceedings of KDD07*, volume 12, pages 210–219, 2007.

[Dai *et al.*, 2007b] W. Dai, Q. Yang, G.R. Xue, and Y. Yu. Boosting for transfer learning. In *Proceedings of the 24th International Conference on Machine learning (ICML)*, pages 193–200. ACM, 2007.

[Dempster *et al.*, 1977] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[Ding *et al.*, 2006] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of KDD06*, 2006.

[Glorot *et al.*, 2011] X. Glorot, A. Bordes, and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine learning (ICML)*, 2011.

[Gupta *et al.*, 2011] S.K. Gupta, D. Phung, B. Adams, and S. Venkatesh. Regularized nonnegative shared subspace learning. *Data Mining and Knowledge Discovery*, pages 1–41, 2011.

[Jin *et al.*, 2011] O. Jin, N.N. Liu, K. Zhao, Y. Yu, and Q. Yang. Transferring topical knowledge from auxiliary long texts for short text clustering. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 775–784. ACM, 2011.

[Lee *et al.*, 1999] D.D. Lee, H.S. Seung, et al. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[Li *et al.*, 2010] T. Li, V. Sindhwani, C. Ding, and Y. Zhang. Bridging domains with words: Opinion analysis with matrix tri-factorizations. In *Proceedings of the 2010 SIAM International Conference on Data Mining (SDM)*, pages 293–302, 2010.

[Liu *et al.*, 2008] Q. Liu, A.J. Mackey, D.S. Roos, and F.C.N. Pereira. Evigan: a hidden variable model for integrating gene evidence for eukaryotic gene prediction. *Bioinformatics*, 24(5):597–605, 2008.

[Long *et al.*, 2012a] M. Long, J. Wang, G. Ding, W. Cheng, X. Zhang, and W. Wang. Dual transfer learning. In *Proceedings of the 12th SIAM International Conference on Data Mining (SDM)*, 2012.

[Long *et al.*, 2012b] M. Long, J. Wang, G. Ding, D. Shen, and Q. Yang. Transfer learning with graph co-regularization. In *Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*, 2012.

[Mansour *et al.*, 2009] Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems (NIPS)*, volume 21, pages 1041–1048, 2009.

[Pan *et al.*, 2008] S.J. Pan, J.T. Kwok, and Q. Yang. Transfer learning via dimensionality reduction. In *Twenty-Third AAAI Conference on Artificial Intelligence (AAAI)*, 2008.

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

[Saenko *et al.*, 2010] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Computer Vision–ECCV 2010*, pages 213–226. Springer, 2010.

[Salton and Buckley, 1988] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[Wang and Mahadevan, 2011] C. Wang and S. Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, volume 2, pages 1541–1546, 2011.

[Xue *et al.*, 2008] G.R. Xue, W. Dai, Q. Yang, and Y. Yu. Topic-bridged plsa for cross-domain text classification. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 627–634. ACM, 2008.

[Zhuang *et al.*, 2011] F. Zhuang, P. Luo, H. Xiong, Q. He, Y. Xiong, and Z. Shi. Exploiting associations between word clusters and document classes for cross-domain text categorization. *Statistical Analysis and Data Mining*, 4(1):100–114, 2011.