Softly Associative Transfer Learning for Cross-Domain Classification

Deqing Wang, *Member, IEEE*, Chenwei Lu, Junjie Wu, Hongfu Liu, Wenjie Zhang, Fuzhen Zhuang, and Hui Zhang

***Abstract*—The main challenge of cross-domain text classification is to train a classifier in a source domain while applying it to a different target domain. Many transfer learning-based algorithms, for example, dual transfer learning, triplex transfer learning, etc., have been proposed for cross-domain classification, by detecting a shared low-dimensional feature representation for both source and target domains. These methods, however, often assume that the word clusters matrix or the clusters association matrix as knowledge transferring bridges are exactly the same across different domains, which is actually unrealistic in real-world applications and, therefore, could degrade classification performance. In light of this, in this paper, we propose a softly associative transfer learning algorithm for cross-domain text classification. Specifically, we integrate two non-negative matrix tri-factorizations into a joint optimization framework, with approximate constraints on both word clusters matrices and clusters association matrices so as to allow proper diversity in knowledge transfer, and with another approximate constraint on class labels in source domains in order to handle noisy labels. An iterative algorithm is then proposed to solve the above problem, with its convergence verified theoretically and empirically. Extensive experimental results on various text datasets demonstrate the effectiveness of our algorithm, even with the presence of abundant state-of-the-art competitors.**

***Index Terms*—Cross-domain text classification, non-negative matrix tri-factorizations (NMTFs), softly associative transfer learning (sa-TL).**

Manuscript received June 28, 2018; revised October 15, 2018; accepted January 2, 2019. Date of publication January 25, 2019; date of current version October 26, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 71501003, Grant 71725002, Grant 71531001, Grant U1636210, Grant U1836206, and Grant 61773361, and in part by the State Key Laboratory of Software Development Environment under Grant SKLSDE-2018ZX-13. This paper was recommended by Associate Editor Y. S. Ong. *(Corresponding author: Junjie Wu.)*

D. Wang, C. Lu, and H. Zhang are with the School of Computer Science, Beihang University, Beijing 100191, China (e-mail: dqwang@buaa.edu.cn).

J. Wu is with the School of Economics and Management, Beihang University, Beijing 100191, China, also with the Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China, and also with the Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations, Beihang University, Beijing 100191, China (e-mail: wujj@buaa.edu.cn).

H. Liu is with the School of Computer Science, Brandeis University, Waltham, MA 02453 USA.

W. Zhang is with the Center of Development and Research, Yidian News Inc., Beijing, China.

F. Zhuang is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2019.2891577

I. Introduction

GIVEN its importance, the cross-domain classification (CDC) problem (also called *domain adaptation*) has attracted much attention from both academia and industries [1]–[13] in recent years. In this problem, the source domain and the target domain share the same space of word features and the same set of class labels, but the distributions of features may vary across domains, which is obviously different from the classical learning problem assuming that training and test instances are from the same distribution. Transfer learning [2], [14], a paradigm aiming to extract knowledge from label-rich source domains to enhance the predictive models of target domains, is an ideal way to address the CDC problem. Many approaches have been proposed for this purpose, for example, co-cluster transfer learning (CoCC) [15], Mtrick [4]; dual knowledge transfer (DKT) [5]; dual transfer learning (DTL) [6]; triplex transfer learning (TriTL) [14], [16]; etc., which achieve knowledge transfer by detecting a shared low-dimensional feature representation for both the source and target domains.

While the above transfer learning-based models have yielded certain progress on CDC tasks, their performance often degrades when the data distributions in the source and target domains differ significantly [17]. This first shows up as the diversity of feature concepts between the source and target domains. For example, in DTL [18] and TriTL [16], the feature clusters matrix is divided into two parts to describe the shared and distinct concepts, respectively. However, they both assume that the distributions of shared concepts are exactly the same for the two domains. In many cases, however, the source and target domains have quite different distributions of word frequency, and the diversity of feature distributions might adversely affect the performance in the target domain by introducing the so-called *negative transferring* [19], [20]. Blitzer *et al.* [21] founded that negative transferring from the domain "Kitchen" to the domain "Book" decreased the performance, and Gui *et al.* [20] suggested detecting negative transferring during knowledge transferring. One possible solution is to allow the diversity of share concepts between two domains. In this paper, we assume that the distributions of feature concepts between two domains are merely similar rather than the same, and we propose an approximate constraint for the shared word concepts.

The other problem rising from different data distributions shows up as the diversity of model knowledge transferred from the source domain to the target one. In previous

studies [4], [16], [18], the models often assume that the whole or part of the concept association matrix is unchanged across different domains, which is actually unreasonable for real-world applications. For example, the target domain is from the DVD reviews, whereas the source domain is from the MUSIC reviews. It is obvious that the concept association matrix for DVD is not the same as that for MUSIC, because users often write different words and topics to evaluate the DVD and MUSIC products. As a consequence, in this paper, we suggest separating a concept association matrix into two parts: 1) the shared part and 2) the distinct one. The knowledge characterized by the shared part between the source and target domains should also be similar rather than exactly the same.

The above problems and ideas indeed motivate this paper. In this paper, we fuse two non-negative matrix tri-factorizations (NMTFs) with softly associative transfer learning (sa-TL) into a joint optimization framework for cross-domain text classification. Our main contributions are summarized as follows.

1) For CDC tasks, we suggest not only taking into account the diversity of feature concepts between the source and the target domains but also incorporating the difference of model knowledge transferred from the source to the target domains. Along this line, we propose an sa-TL method, in which soft regularization constraints for both word clusters matrices and clusters association matrices are added to allow diversity in knowledge transfer.
2) We introduce an extension of categorical utility function into our sa-TL model so as to loose the constraint of label matrix. This empowers the sa-TL model with the ability in dealing with label noise in the source domain.
3) We integrate two NMTFs into a joint optimization function and derive an efficient iterative optimization algorithm. We also provide theoretical proof and empirical verification to its convergence, then we analyze the computational complexity of the algorithm.
4) We conduct extensive experiments on multilingual Amazon product reviews and 20 Newsgroup datasets with abundant state-of-the-art methods as competitors. The results demonstrate the advantage of the sa-TL model stemming from sa-TL.

The remainder of this paper is organized as follows. Section II reviews some related work on CDC. We then introduce some preliminary knowledge on matrix factorization in Section III. We formally present our model in Section IV and give the experimental setup and results in Sections V and VI, respectively. Finally, Section VII summarizes this paper.

II. Related Work

In this section, we briefly review the recent studies on CDC from the transfer learning perspective.

For the CDC problem, the key idea behind it is to discover the knowledge bridge between the source domain and the target domain. Therefore, some studies exploited the commonality between different domains for knowledge transfer. For example, Dai *et al.* [15] proposed a co-clustering-based method (named CoCC), which identified the word clusters across different domains by propagating the class information and knowledge from the source domain to the target domain. CoCC, however, only considers the identical concepts across different domain and results in a performance decrease for cross-domain tasks. Then, Li *et al.* [22] proposed sharing the same word clusters between the source and target domains to transfer label information. However, the word clusters between the source and target domains are only related, rather than exactly the same.

Further, recent studies argued that the high-level concepts help to model the difference of data distributions and are more appropriate for text categorization tasks. Specifically, these methods assume that the feature space across different domains can be divided into the same/similar and the distinct parts, that is, the shared or identical concepts, alike concepts, and distinct concepts, which are used as the bridge for knowledge transfer. For instance, DKT [5] was the first attempt to discover the alike concepts and to use them for knowledge transfer. Then, DTL [18] modeled the shared concepts by distinguishing the identical and alike concepts to establish the bridge from the source domain to the target domain. Tri-TL [16] and HIDC [23] further exploited the identical concepts, alike concepts, and distinct concepts to train the predictive model based on the shared and alike concepts. Zhuang *et al.* [14] improved Tri-TL by adding a regularized manifold structure of the target domain. Wang and Yang [24] presented a novel algorithm to find the structural similarity between two domains to enable transfer learning at a structured knowledge level. For the multisource domain problem, Hu *et al.* [25] proposed the multiknowledge transfer from multisource domains to one target domain. Besides direct knowledge transfer, transitive transfer learning (TTL) [26] was also proposed, which aimed at breaking the large domain distances and transferring knowledge even when the source and target domains share few factors directly. In TTL, the source and target domains can be connected by intermediate domains through some underlying factors. In these studies, the high-level concepts are utilized with the observation that different domains may use different key words to express the same concepts.

Another research direction is how to transfer the model across different domains. For example, Evgeniou and Pontil [27] borrowed the idea of the Bayesian framework to SVMs for multitask learning and assumed that the parameter w in SVMs for each task can be separated into two terms: one is a common term over tasks and the other is a task-specific term. Gao *et al.* [1] proposed a locally weighted ensemble learning framework to combine multiple models for transfer learning. Zhuang *et al.* [4] exploited the assumption of an unchanged association matrix between the word features concepts and the example classes as the bridge across different domains. Long *et al.* [28] introduced a novel transfer learning method with graph co-regularization, in which the graph properties between examples in class interior and exterior were both used as regularization constraints.

To summarize, the abovementioned transfer learning approaches achieve better performance than nontransferred methods for cross-domain text classification tasks. However, they are two-stages transfer learning and could not consider

TABLE I
MATH NOTATION

$\mathcal{D}$	the data set
X	the word-document co-occurrence matrix from a domain
n	the number of documents
m	the dimensionality of features
c	the number of document classes
n_s	the number of documents in the source domain
n_t	the number of documents in the target domain
k	the number of feature clusters, $k = k_1 + k_2$
k_1	the number of similar feature clusters
k_2	the number of distinct feature clusters
F	the matrix of feature clusters
H	the matrix of the association between feature clusters and document classes
G	the matrix of the document classes
M	the alignment matrix of document classes

transfer learning as a whole framework. For example, DTL and TriTL focus on knowledge transfer (assuming feature clusters matrix is the bridge between two domains), whereas Mtrick focuses on model transfer (assuming association matrix is the unchanged bridge between two domains). This is the first drawback of the existing methods. The next drawback is that the assumption of unchanged knowledge or model transfer (part or whole) of the existing methods is too strict for cross-domain learning tasks, which affect the improvement of performance. In this paper, we will propose our method by considering the two sides together with softly associative constraints. Moreover, we also reduce the adverse impact of label noise which occurs in the source domain by adopting an extension of categorical utility function.

III. PRELIMINARIES

In this section, we first introduce some mathematical notations. Data matrices are represented by uppercase, such as X and Y, in which, $X_{(ij)}$ is the ith row and jth column element. Datasets are denoted by Calligraphic letters, such as $\mathcal{D}$. Bold letters, such as $\mathbf{u}$, are used to represent vectors. We use $\mathbb{R}$ and $\mathbb{R}_+$ to represent the set of real numbers and non-negative real numbers, respectively. Table I shows the details of notations and their denotations.

Non-negative matrix factorization (NMF) [29], [30] is a popular and effective technique for data clustering, classification, pattern recognition, etc. In NMF, a term-document matrix $X \in {\mathbb{R}_+}^{m\times n}$ is decomposed into a product of two factor matrices $F \in {\mathbb{R}_+}^{m\times k}$ and $G^\top \in {\mathbb{R}_+}^{n\times k}$, where F corresponds to cluster centers and G is associated with cluster indicator vectors. Generally, an orthogonality constraint is often imposed on a factor matrix in the decomposition and provides a clearer interpretation on a link between clustering and matrix decomposition [31]. However, orthogonal NMF often gives a rather poor matrix low-rank approximation because of its restrictive orthogonality. Thus, Ding *et al.* [32] proposed an NMTF method, in which an extra factor H representing how document clusters are related to feature clusters was introduced to absorb the different scales of X, F, and G. Compared to two-factor decomposition, the term-document matrix X is decomposed into three submatrices, which can be factorized by solving the following optimization problem:

$$\min_{F,H,G\geq 0} \left|\left|X - FHG^\top\right|\right|^2 \tag{1}$$

where $||\cdot||$ denotes the Frobenius norm of the matrix, and $F \in {\mathbb{R}_+}^{m\times k}$, $H \in {\mathbb{R}_+}^{k\times c}$, and $G \in {\mathbb{R}_+}^{n\times c}$ are the feature clusters matrix, association matrix, and document classes matrix, respectively.

Generally, NMTF gives a good framework for simultaneously clustering the rows and columns of X, and the factorized matrices (F, H, G) have their own semantic meanings as follows [32].

1) $F = [\mathbf{f_1},\ldots,\mathbf{f_k}]$ is an $m \times k$ cluster assignment matrix representing feature clusters, where $\mathbf{f_i}$ is a probability distribution over m features and is referred to as a feature cluster.
2) G is an $n \times c$ document clusters matrix, where $G_{(ij)}$ is the probability that the ith document belongs to the jth document cluster. For the classification task, each document cluster is considered as a class or category.
3) $H = [\mathbf{h_1},\ldots,\mathbf{h_c}]$ is a $k \times c$ weight matrix representing the association between the word clusters and document clusters. $H_{(ij)}$ is the probability (weights) that the ith feature cluster is associated with the jth document cluster.

For CDC tasks, given a labeled source domain dataset $\mathcal{D}_s$ and an unlabeled target domain dataset $\mathcal{D}_t$, then we have $\mathcal{D} = (\mathcal{D}_1, \ldots, \mathcal{D}_{n_s}, \mathcal{D}_{n_s+1}, \ldots, \mathcal{D}_{n_s+n_t})$, where n_s and n_t are the numbers of documents in the source domain and in the target domain, respectively. Let $X = (\mathbf{x}_1, \ldots, \mathbf{x}_{n_s}, \mathbf{x}_{n_s+1}, \ldots, \mathbf{x}_{n_s+n_t})$ be the corresponding word-document co-occurrence matrix of $\mathcal{D}$, and each example $\mathbf{x_i}$ $(1 \leq i \leq n_s)$ is associated with one of the c class labels, that is, $\mathbf{y_{ij}} = 1$ if $\mathbf{x_i}$ belongs to class j $(1 \leq j \leq c)$. The purpose of transfer learning is to improve the predictive ability on the target domain $\mathcal{D}_t$ by using transferred knowledge learned from source domain $\mathcal{D}_s$.

Recently, some state-of-the-art approaches based on NMTF have been proposed [4], [14], [16], [18], [25], [33] to solve the above cross-domain text classification tasks. Generally, the framework of these NMTF-based transfer learning approaches is often formulated as

$$\min_{F_s,G_s,H,F_t,G_t\geq 0} \left|\left|X_s - F_sHG_s^\top\right|\right|^2 + \left|\left|X_t - F_tHG_t^\top\right|\right|^2 \tag{2}$$

where X, F, G, and H have the same definitions as in (1), and the subscripts s and t denote the source domain and the target domain, respectively.

IV. MODEL AND ALGORITHM

In this section, we present the model formulation and the iterative optimization algorithm of our proposed sa-TL, then we analyze the complexity of sa-TL.

A. Model Formulation

Generally, the existing framework of NMTF-based transfer learning has three shortcomings: 1) ignoring the diversity of feature clusters (concepts) across different domains; 2) imposing the strict assumption on knowledge and/or model transfer

from the source domain to the target domain; and 3) decreasing the performance severely when noisy labels occur in the source domain. For CDC tasks, the documents are from related but different domains; thus, they own different features distributions, even though they share the same feature space. Moreover, the bridge of knowledge transfer across different domains, that is, the unchanged association matrix ($H_s = H_t = H$), is too strict for real-world CDC tasks. Finally, the previous models completely ignore the label noise in the source domain and result in the serious decline of predictive performance in the target domain.

In this paper, we propose an sa-TL algorithm, which loosens the strict constraint of both word clusters matrices and clusters association matrices, and only assumes that the difference of word clusters matrices and clusters association matrices between the source and target domains is approximately smaller, rather than being kept unchanged. Specifically, taking the source domain D_s as an example, in sa-TL, we separate the feature clusters matrix $F_s \in {\mathbb{R}_+}^{m\times k}$ into a similar part $F_{ss} \in {\mathbb{R}_+}^{m\times k_1}$ and a distinct part $F_{sd} \in {\mathbb{R}_+}^{m\times k_2}$. Thus, the corresponding $H_s \in {\mathbb{R}_+}^{k\times c}$ also has two parts: $H_{ss} \in {\mathbb{R}_+}^{k_1\times c}$ and $H_{ts} \in {\mathbb{R}_+}^{k_2\times c}$. Note that k_1 and k_2 are the numbers of similar feature clusters and distinct ones, respectively. We also separate F_t and H_t in the target domain as the above way.

Then, for similar feature clusters, we employ an L_2 constraint to guarantee the approximation and the smooth of word distributions across different domains. And an L_1 constraint is adopted for the distinct concepts to guarantee the sparseness of word distributions, which are specific to different domains. For the same reason, an L_2 constraint factor ensures that the corresponding knowledge transfer from the source domain to the target domain is not only approximately similar but also frequently appears between two domains; an L_1 constraint factor is added to ensure the sparseness of the specific knowledge across different domains. Moreover, we also introduce an extension of the categorical utility function into our sa-TL model to adapt the label noise in the source domain by loosening the constraint of the label matrix. Concretely, we use the ground truth of source label G_0 as the supervised information by requiring that G_s be similar to G_0. Finally, we integrate the two NMTFs into one formulation and propose the joint optimization formulation, as shown in

$$\begin{aligned}
\min_{\substack{F_{ss},F_{sd},H_{ss},H_{sd},G_s\\ F_{ts},F_{td},H_{ts},H_{td},G_t\geq 0}} & ||X_s - \begin{bmatrix}F_{ss} & F_{sd}\end{bmatrix}\begin{bmatrix}H_{ss}\\ H_{sd}\end{bmatrix}G_s^\top||^2 \\
& + ||X_t - \begin{bmatrix}F_{ts} & F_{td}\end{bmatrix}\begin{bmatrix}H_{ts}\\ H_{td}\end{bmatrix}G_t^\top||^2 + \pi||G_0 - G_sM||^2 \\
& + \alpha||F_{ss} - F_{ts}||^2 + \beta||H_{ss} - H_{ts}||^2 \\
& + \gamma||F_{sd} + F_{td}|| + \delta||H_{sd} + H_{td}|| \\
\text{s.t.} & \sum_{i=1}^{m} F_{ss_{(ij)}} = 1,\ \sum_{i=1}^{m} F_{ts_{(ij)}} = 1, \sum_{i=1}^{m} F_{sd_{(ij)}} = 1 \\
& \sum_{i=1}^{m} F_{td_{(ij)}} = 1, \sum_{j=1}^{c} G_{s_{(ij)}} = 1, \sum_{j=1}^{c} G_{t_{(ij)}} = 1
\end{aligned} \tag{3}$$

where $\alpha, \beta, \gamma, \delta, \pi \geq 0$ are the tradeoff parameters and $M \in {\mathbb{R}_+}^{c\times c}$ is an alignment matrix between document clusters and document classes. It is worth noting that the third term $\pi||G_0 - G_sM||^2$ in objective function in (3) works as a regularizer to measure the similarity between the ground truth of source label G_0 and the learned G_s in the partition level. Since the clustering is orderless, M is introduced to learn the best mapping between G_0 and G_s. Actually, the third term can be regarded as the categorical utility function $U_c(G_0, G_s)$ in consensus clustering [34], where the equivalent relationship among the categorical utility function and the third term is uncovered in our previous work [35]. The difference is that the input of the categorical utility function requires the binary matrices, while G_s is the continuous indictor matrix in our objective function, which can be regarded as an extension of the categorical utility function. After minimizing (3), the class label of the ith instance in the target domain D_t is the index of maximum value in $G_{t_{(i,:)}}$.

Remark: Alone this line, we can resurvey the previous transfer learning methods according to the framework in (3) and point out their shortcomings. The model in [22] adopted the shared word clusters between the source and target domains to transfer label information. However, the word clusters are only related, rather than exactly the same; thus, this assumption will cause incorrect knowledge transfer. Moreover, model transfer is totally ignored. Mtrick [4] considered the diversity of word clusters and exploited the unchange associations as a bridge of knowledge between two domains, which may cause incorrect model transfer because the association matrices could be slightly different, accounting for the domain variations. Further, DTL [18] considered both the diversity of word clusters and the common model knowledge, but it still assumed that the common association was unchanged and the specific association was not explicitly combined into an objective function. TriTL [16] further divided word clusters into three parts (identical, alike, and distinct) and the corresponding association was also divided into three parts. Although TriTL manually distinguished between different concepts and the association matrix, it still kept the association matrix unchanged which will cause incorrect model factorization and knowledge transfer. Moreover, TriTL does not consider the label noise in the source domain.

B. Model Inference

The solution of the optimization problem in (3) is not concave and it is difficult to acquire the global solution. Therefore, we derive an iterative optimization algorithm to search for the local optimal solution. For the minimization problem of (3), it can be transformed into the minimization problem of (4) by applying the relation between the Frobenius norm of the matrix and trace of the matrix

$$\begin{aligned}
\min_{\substack{F_{ss},F_{sd},H_{ss},H_{sd},G_s,\\ F_{ts},F_{td},H_{ts},H_{td},G_t\geq 0}} & \mathcal{L} \\
= & \operatorname{Tr}(X_s^\top X_s - 2X_s^\top T_s G_s^\top + G_s T_s^\top T_s G_s^\top) \\
& + \operatorname{Tr}(X_t^\top X_t - 2X_t^\top T_t G_t^\top + G_t T_t^\top T_t G_t^\top) \\
& + \pi \cdot \operatorname{Tr}(G_0^\top G_0 - 2G_0^\top G_s M + M^\top G_s^\top G_s M) \\
& + \alpha \cdot \operatorname{Tr}(F_{ss}^\top F_{ss} - 2F_{ss}^\top F_{ts} + F_{ts}^\top F_{ts}) \\
& + \beta \cdot \operatorname{Tr}(H_{ss}^\top H_{ss} - 2H_{ss}^\top H_{ts} + H_{ts}^\top H_{ts})
\end{aligned}$$

$$+\ \gamma \cdot \sum_{i=1}^{m}\sum_{j=1}^{k_2}|F_{sd_{(ij)}}| + \gamma \cdot \sum_{i=1}^{m}\sum_{j=1}^{k_2}|F_{td_{(ij)}}| + \delta \cdot \sum_{i=1}^{k_2}\sum_{j=1}^{c}|H_{sd_{(ij)}}| + \delta \cdot \sum_{i=1}^{k_2}\sum_{j=1}^{c}|H_{td_{(ij)}}|$$
$$\text{s.t.}\ \sum_{i=1}^{m}F_{ss_{(ij)}} = 1,\ \sum_{i=1}^{m}F_{ts_{(ij)}} = 1, \sum_{i=1}^{m}F_{sd_{(ij)}} = 1$$
$$\sum_{i=1}^{m}F_{td_{(ij)}} = 1, \sum_{j=1}^{c}G_{s_{(ij)}} = 1, \sum_{j=1}^{c}G_{t_{(ij)}} = 1 \tag{4}$$

where $T_s = F_{ss}H_{ss} + F_{sd}H_{sd}$ and $T_t = F_{ts}H_{ts} + F_{td}H_{td}$.

Then, we obtain the partial differential of (4), shown as

$$\frac{\partial L}{\partial F_{ss}} = 2T_sG_s^\top G_sH_{ss}^\top + 2\alpha \cdot (F_{ss} - F_{ts}) - 2X_sG_sH_{ss}^\top$$
$$\frac{\partial L}{\partial H_{ss}} = 2F_{ss}^\top T_sG_s^\top G_s + 2\beta \cdot (H_{ss} - H_{ts}) - 2F_{ss}^\top X_sG_s$$
$$\frac{\partial L}{\partial F_{sd}} = 2F_{sd}H_{sd}G_s^\top G_sH_{sd}^\top + \gamma - 2(X_s - F_{ss}H_{ss}G_s^\top)G_sH_{sd}^\top$$
$$\frac{\partial L}{\partial H_{sd}} = 2F_{sd}^\top F_{sd}H_{sd}G_s^\top G_s + \delta - 2F_{sd}^\top(X_s - F_{ss}H_{ss}G_s^\top)G_s$$
$$\frac{\partial L}{\partial G_s} = 2G_sT_s^\top T_s - 2X_s^\top T_s - 2\pi \cdot G_0M^\top + 2\pi \cdot G_sMM^\top$$
$$\frac{\partial L}{\partial M} = \pi \cdot (-2G_s^\top G_0 + 2G_s^\top G_sM)$$
$$\frac{\partial L}{\partial F_{ts}} = 2T_tG_t^\top G_tH_{ts}^\top + 2\alpha \cdot (F_{ts} - F_{ss}) - 2X_tG_tH_{ts}^\top$$
$$\frac{\partial L}{\partial H_{ts}} = 2F_{ts}^\top T_tG_t^\top G_t + 2\beta \cdot (H_{ts} - H_{ss}) - 2F_{ts}^\top X_tG_t$$
$$\frac{\partial L}{\partial F_{td}} = 2F_{td}H_{td}G_t^\top G_tH_{td}^\top + \gamma - 2(X_t - F_{ts}H_{ts}G_t^\top)G_tH_{td}^\top$$
$$\frac{\partial L}{\partial H_{td}} = 2F_{td}^\top F_{td}H_{td}G_t^\top G_t + \delta - 2F_{td}^\top(X_t - F_{ts}H_{ts}G_t^\top)G_t$$
$$\frac{\partial L}{\partial G_t} = 2G_tT_t^\top T_t - 2X_t^\top T_t. \tag{5}$$

From the partial differential in (5), we employ the Lagrangian function and the auxiliary function [32] to deduce the following iterative updating rules (6)–(16). Then we develop an alternately iterative algorithm, which can converge to a local optimal solution. These matrices are updated in each round of iteration as follows and the detailed deduced and theoretical proof of convergence are presented in the following section:

$$F_{ss_{(ij)}} \leftarrow F_{ss_{(ij)}} \cdot \sqrt{\frac{(X_sG_sH_{ss}^\top + \alpha \cdot F_{ts})_{(ij)}}{(T_sG_s^\top G_sH_{ss}^\top + \alpha \cdot F_{ss})_{(ij)}}} \tag{6}$$
$$H_{ss_{(ij)}} \leftarrow H_{ss_{(ij)}} \cdot \sqrt{\frac{(F_{ss}^\top X_sG_s + \beta \cdot H_{ts})_{(ij)}}{(F_{ss}^\top T_sG_s^\top G_s + \beta \cdot H_{ss})_{(ij)}}} \tag{7}$$
$$F_{sd_{(ij)}} \leftarrow F_{sd_{(ij)}} \cdot \sqrt{\frac{(2(X_s - F_{ss}H_{ss}G_s^\top)G_sH_{sd}^\top)_{(ij)}}{(2F_{sd}H_{sd}G_s^\top G_sH_{sd}^\top + \gamma)_{(ij)}}} \tag{8}$$
$$H_{sd_{(ij)}} \leftarrow H_{sd_{(ij)}} \cdot \sqrt{\frac{(2F_{sd}^\top(X_s - F_{ss}H_{ss}G_s^\top)G_s)_{(ij)}}{(2F_{sd}^\top F_{sd}H_{sd}G_s^\top G_s + \delta)_{(ij)}}} \tag{9}$$
$$G_{s_{(ij)}} \leftarrow G_{s_{(ij)}} \cdot \sqrt{\frac{(X_s^\top T_s + \pi \cdot G_0M^\top)_{(ij)}}{(G_sT_s^\top T_s + \pi \cdot G_sMM^\top)_{(ij)}}} \tag{10}$$
$$M_{(ij)} \leftarrow M_{(ij)} \cdot \sqrt{\frac{(G_s^\top G_0)_{ij}}{(G_s^\top G_s)_{ij}}} \tag{11}$$
$$F_{ts_{(ij)}} \leftarrow F_{ts_{(ij)}} \cdot \sqrt{\frac{(X_tG_tH_{ts}^\top + \alpha \cdot F_{ss})_{(ij)}}{(T_tG_t^\top G_tH_{ts}^\top + \alpha \cdot F_{ts})_{(ij)}}} \tag{12}$$
$$H_{ts_{(ij)}} \leftarrow H_{ts_{(ij)}} \cdot \sqrt{\frac{(F_{ts}^\top X_tG_t + \beta \cdot H_{ss})_{(ij)}}{(F_{ts}^\top T_tG_t^\top G_t + \beta \cdot H_{ts})_{(ij)}}} \tag{13}$$
$$F_{td_{(ij)}} \leftarrow F_{td_{(ij)}} \cdot \sqrt{\frac{(2(X_t - F_{ts}H_{ts}G_t^\top)G_tH_{td}^\top)_{(ij)}}{(2F_{td}H_{td}G_t^\top G_tH_{td}^\top + \gamma)_{(ij)}}} \tag{14}$$
$$H_{td_{(ij)}} \leftarrow H_{td_{(ij)}} \cdot \sqrt{\frac{(2F_{td}^\top(X_t - F_{ts}H_{ts}G_t^\top)G_t)_{(ij)}}{(2F_{td}^\top F_{td}H_{td}G_t^\top G_t + \delta)_{(ij)}}} \tag{15}$$
$$G_{t_{(ij)}} \leftarrow G_{t_{(ij)}} \cdot \sqrt{\frac{(X_t^\top T_t)_{(ij)}}{(G_tT_t^\top T_t)_{(ij)}}}. \tag{16}$$

Here, G_0 is unchanged during the iteration steps, because it is the ground truth of the source domain data. As shown in (17), F_{sd}, F_{td}, F_{ss}, and F_{ts} are column normalized and G_s and G_t are the row normalized in order to satisfy the equality constraints of the objective function. We provide the experimental verification of convergence of the iterative rules in the next section

$$F_{ss_{(\cdot j)}} \leftarrow \frac{F_{ss_{(\cdot j)}}}{\sum F_{ss_{(\cdot j)}}}, F_{sd_{(\cdot j)}} \leftarrow \frac{F_{sd_{(\cdot j)}}}{\sum F_{sd_{(\cdot j)}}}$$
$$F_{ts_{(\cdot j)}} \leftarrow \frac{F_{ts_{(\cdot j)}}}{\sum F_{ts_{(\cdot j)}}}, F_{td_{(\cdot j)}} \leftarrow \frac{F_{td_{(\cdot j)}}}{\sum F_{td(\cdot j)}}$$
$$G_{s_{(i\cdot)}} \leftarrow \frac{G_{s_{(i\cdot)}}}{\sum G_{s_{(i\cdot)}}}, G_{t_{(i\cdot)}} \leftarrow \frac{G_{t_{(i\cdot)}}}{\sum G_{t_{(i\cdot)}}}. \tag{17}$$

C. Analysis of Convergence

Here, we present the theoretic analysis of algorithm convergence by using our updating rules, that is, (6)–(16). Taking F_{sd} as an example, we first show how to deduce (8) and check the convergence of F_{sd} when the remaining parameters are fixed. According to the minimization problem of (4) and omitting the items which are independent of F_{sd}, we formulate the optimization problem with constraints as the following Lagrangian function:

$$\mathcal{G}(F_{sd}) = \text{Tr}(-2(X_s - F_{ss}H_{ss}G_s^\top)G_sF_{sd}H_{sd} + G_sH_{sd}^\top F_{sd}^\top F_{sd}H_{sd}G_s^\top + \gamma \cdot \sum_{i=1}^{m}\sum_{j=1}^{k_2}|F_{sd_{(ij)}}| + \text{Tr}(\boldsymbol{\lambda}(F_{sd}^\top \mathbf{1}_m - \mathbf{1}_{k_2})(F_{sd}^\top \mathbf{1}_m - \mathbf{1}_{k_2})^\top) \tag{18}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^{k_2 \times k_2}$ is a diagonal matrix.

Then, we can compute the differential of (18) with respect to F_{sd} as follows:

$$\frac{\partial \mathcal{G}}{\partial F_{sd}} = 2F_{sd}H_{sd}G_s^\top G_sH_{sd}^\top + \gamma + 2 \cdot \mathbf{1}_m\mathbf{1}_m{}^\top F_{sd}\boldsymbol{\lambda} - 2(X_s - F_{ss}H_{ss}G_s^\top)G_sH_{sd}^\top - 2 \cdot \mathbf{1}_m\mathbf{1}_{k_2}{}^\top)\boldsymbol{\lambda}. \tag{19}$$

Algorithm 1 sa-TL for CDC

Input:

the source domain $\mathcal{D}_s$ and its corresponding word-document matrix $X_s \in \mathbb{R}_+^{m\times n_s}$;
the label vector of the i-th instance is $\mathbf{y}_i$ with $y_{ij} = 1$ if it belongs to class j.
the target domain $\mathcal{D}_t$ and its corresponding word-document matrix $X_t \in \mathbb{R}_+^{m\times n_t}$;
α,β,γ, δ, π, the maximal number of iteration max_{iter}, and the error threshold ε;
k_1: the number of similar feature clusters; k_2: the number of distinct feature clusters;

Output:

$F_{ss}, F_{sd}, H_{ss}, H_{sd}, G_s, M, F_{ts}, F_{td}, H_{ts}, H_{td}, G_t$.

1. Initialize the matrix F_{ss}, F_{sd}, H_{ss}, H_{sd}, G_s, M, F_{ts}, F_{td}, H_{ts}, H_{td}, G_t;
2. Calculate the value of $\mathcal{L}^{(0)}$ *of Eq.* 3;
3. $iter \coloneqq 1$;
4. while ($iter \le max_{iter}$) {
5. Update $F_{ss}^{(iter)}$ according to Eq. 6 and normalize it based on Eq. 17;
6. Update $H_{ss}^{(iter)}$ according to Eq. 7;
7. Update $F_{sd}^{(iter)}$ according to Eq. 8 and normalize it based on Eq. 17;
8. Update $H_{sd}^{(iter)}$ according to Eq. 9;
9. Update $G_s^{(iter)}$ according to Eq. 10 and normalize it based on Eq. 17;
10. Update $M^{(iter)}$ according to Eq. 11;
11. Update $F_{ts}^{(iter)}$ according to Eq. 12 and normalize it based on Eq. 17;
12. Update $H_{ts}^{(iter)}$ according to Eq. 13;
13. Update $F_{td}^{(iter)}$ according to Eq. 14 and normalize it based on Eq. 17;
14. Update $H_{td}^{(iter)}$ according to Eq. 15;
15. Update $G_t^{(iter)}$ according to Eq. 16 and normalize it based on Eq. 17;
16. Calculate $\mathcal{L}^{(iter)}$. if $|\mathcal{L}^{(iter)} - \mathcal{L}^{(iter-1)}| < \varepsilon$, turn to step 18;
17. $iter \coloneqq iter + 1$;
18. }
19. Output $F_{ss}^{(iter)}$, $F_{sd}^{(iter)}$, $H_{ss}^{(iter)}$, $H_{sd}^{(iter)}$, $G_s^{(iter)}$, $M^{(iter)}$ $F_{ts}^{(iter)}$, $F_{td}^{(iter)}$, $H_{ts}^{(iter)}$, $H_{td}^{(iter)}$, $G_t^{(iter)}$;

Letting $(\partial\mathcal{G}/\partial F_{sd}) = 0$, we can obtain the updating rule in

$$F_{sd_{(ij)}} \leftarrow F_{sd_{(ij)}} \cdot \sqrt{\frac{(2(X_s - F_{ss}H_{ss}G_s^\top)G_sH_{sd}^\top + \mathbf{1}_m\mathbf{1}_{k_2}^\top\boldsymbol{\lambda})_{(ij)}}{(2(F_{sd}H_{sd}G_s^\top G_sH_{sd}^\top + \mathbf{1}_m\mathbf{1}_m^\top F_{sd}\boldsymbol{\lambda}) + \gamma)_{(ij)}}}. \tag{20}$$

Then, we have the following lemma.

Lemma 1: Using the updating rule in (18), (20) will monotonously decrease.

Proof: To prove Lemma 1, we first define an auxiliary function [32] as follows.

Definition 1 (Auxiliary Function): A function $\mathcal{Q}(Y, \widetilde{Y})$ is called an auxiliary function of $\mathcal{T}(Y)$ if it satisfies

$$\mathcal{Q}(Y, \widetilde{Y}) \ge \mathcal{T}(Y), \mathcal{Q}(Y, Y) = \mathcal{T}(Y) \tag{21}$$

for any Y, $\widetilde{Y}$.

Then, define

$$Y^{(t+1)} = \arg\min_Y \mathcal{Q}(Y, Y^{(t)}). \tag{22}$$

Through this definition

$$\mathcal{T}(Y^{(t)}) = \mathcal{Q}(Y^{(t)}, Y^{(t)}) \ge \mathcal{Q}(Y^{(t+1)}, Y^{(t)}) \ge \mathcal{T}(Y^{(t+1)}).$$

It means that minimizing the auxiliary function of $\mathcal{Q}(Y, Y^{(t)})$ ($Y^{(t)}$ is fixed) has the effect to decrease the function of $\mathcal{T}$.

Now, we can construct the auxiliary function of $\mathcal{G}$ as

$$\begin{aligned}
&\mathcal{Q}(F_{sd}, F_{sd}{}') \\
&= \sum_{i=1}^{m}\sum_{j=1}^{k_2}\Bigg\{ -2\cdot\left(X_s - F_{ss}H_{ss}G_s^\top G_sH_{sd}^\top\right)_{[i,j]}F_{sd}{}'_{[i,j]}\left(1+\log\frac{F_{sd[i,j]}}{F_{sd}{}'_{[i,j]}}\right) \\
&\qquad -2\cdot\left(\mathbf{1}_m\mathbf{1}_{k_2}{}^\top\boldsymbol{\lambda}\right)_{[i,j]}F_{sd}{}'_{[i,j]}\left(1+\log\frac{F_{sd[i,j]}}{F_{sd}{}'_{[i,j]}}\right) \\
&\qquad +\left(G_sH_{sd}^\top F'_{sd}H_{sd}G_s^\top + \mathbf{1}_m\mathbf{1}_m{}^\top F_{sd}{}'\boldsymbol{\lambda}\right)_{[i,j]}\frac{F_{sd[i,j]}F_{sd[i,j]}}{F_{sd}{}'_{[i,j]}} \\
&\qquad +\gamma F'_{sd}\left(1+\log\frac{F_{sd[i,j]}}{F_{sd}{}'_{[i,j]}}\right)\Bigg\}.
\end{aligned}$$

Obviously, when $F_{sd} = F_{sd}{}'$, the equality $\mathcal{Q}(F_{sd}, F_{sd}{}') = \mathcal{G}(F_{sd})$ holds. Meanwhile, we can prove the inequality $\mathcal{Q}(F_{sd}, F_{sd}{}') \ge \mathcal{G}(F_{sd})$ holds by using the similar proof approach in [32]. Then, while fixing $F_{sd}{}'$, we minimize $\mathcal{Q}(F_{sd}, F_{sd}{}')$. The differential of $\mathcal{Q}(F_{sd}, F_{sd}{}')$ is

$$\begin{aligned}
&\frac{\partial\mathcal{Q}(F_{sd}, F_{sd}{}')}{\partial F_{sd[i,j]}} \\
&= -2\cdot\left(X_s - F_{ss}H_{ss}G_s^\top G_sH_{sd}^\top\right)_{[i,j]}\frac{F_{sd}{}'_{[i,j]}}{F_{sd[i,j]}} - 2\cdot\left(\mathbf{1}_m\mathbf{1}_{k_2}{}^\top\boldsymbol{\lambda}\right)_{[i,j]}\frac{F_{sd}{}'_{[i,j]}}{F_{sd[i,j]}} \\
&\quad +2\cdot\left(G_sH_{sd}^\top F'_{sd}H_{sd}G_s^\top + \mathbf{1}_m\mathbf{1}_m{}^\top F_{sd}{}'\boldsymbol{\lambda}\right)_{[i,j]}\frac{F_{sd[i,j]}}{F_{sd}{}'_{[i,j]}} + \gamma\frac{F_{sd}{}'_{[i,j]}}{F_{sd[i,j]}}.
\end{aligned}$$

Letting $([\partial \mathcal{Q}(F_{sd}, F_{sd}')]/\partial F_{sd[i,j]}) = 0$, we can obtain (20). Thus, the updating rule in (20) decreases the values of $\mathcal{G}(F_{sd})$. Then, Lemma 1 holds. ■

Compared with (8), (20) contains one more term about $\boldsymbol{\lambda}$, in which $\boldsymbol{\lambda}$ is to drive the solution to satisfy the constrained condition, that is, $\sum_{i=1}^{m} F_{sd_{(ij)}} = 1$. Here, we adopt the normalization technology in [4] to satisfy the constrains regardless of $\boldsymbol{\lambda}$. Specifically, in each iteration, we use (17) to normalize F_{sd}. After normalization, $\mathbf{1}_m \mathbf{1}_{k_2}{}^{\top} \boldsymbol{\lambda}$ is equal to $\mathbf{1}_m \mathbf{1}_m{}^{\top} F_{sd} \boldsymbol{\lambda}$ which are both constants. Therefore, the effect of (8) and (17) can be approximately equivalent to (20) when only considering the convergence. In our solution, we adopt the approximate updating rule of (8) by omitting the items which depend on $\boldsymbol{\lambda}$ in (20). We can use the similar strategy to analyze the convergence of the updating rules in (6)–(16), respectively.

Theorem 1 (Convergence): After each round of iteration in Algorithm 1, the objective function in (4) will not increase.

According to the lemmas for the convergence analysis on the updating rules from (6) to (16), and the multiplicative update rules [30], each updating step in Algorithm 1 will not increase (4) and the objective has a lower bounded by zero, which guarantees the convergence. Thus, the above theorem holds.

D. Algorithm and Computational Complexity

The algorithm based on the above mathematical induction is simple to implement, and the pseudocode of sa-TL is shown in Algorithm 1. After the initialization of parameters, we update $F_{ss}, F_{sd}, H_{ss}, H_{sd}, G_s, M, F_{ts}, F_{td}, H_{ts}, H_{td}, G_t$ according to (6)–(16), and we normalize $F_{ss}, F_{sd}, G_s, F_{ts}, F_{td}, G_t$ based on (17) in each iteration step. The iterative updating ends until reaching the maximal iterative time or the convergence threshold.

Here, we analyze the computational complexity of our proposed sa-TL method from each round of iteration in Algorithm 1. In each iteration step, we have 11 matrices to update. Here, we take F_{ss} as an example to calculate its computational complexity. As shown in (6), the computation of F_{ss} is mainly about the multiplication operation; thus, the complexity of updating F_{ss} is $O(3mk_1c + mk_2c + 3mcn_s)$. Then, we can calculate the computational complexity of $F_{sd}, H_{ss}, H_{sd}, G_s, M$ in the source domain and that of $F_{ts}, F_{td}, H_{ts}, H_{td}, G_t$ in the target domain. Generally, we have $c \ll m$, $k \ll m$, $c \ll n_s/n_t$, and $k \ll n_s/n_t$; thus, the complexity of each iteration step can be simplified as $O(mn)$, where $n = n_s + n_t$. Because we have $\max_{\text{iter}}$ iteration steps, the maximal computational complexity of Algorithm 1 is $O(\max_{\text{iter}} mn)$.

V. Experimental Setup

In this section, we will introduce the common datasets, baseline methods, and experimental setup.

A. Datasets

Two common cross-domain datasets are used in our experiments, that is, Amazon product reviews and 20 Newsgroup. First, we conduct a set of experiments on the well-known multilingual Amazon product reviews[1] used a lot in previous research [2], [4], [6], [16], [36]–[38], which consists of three product categories: 1) BOOKS (B); 2) DVD (D); and 3) MUSIC (M). It has four languages reviews: 1) English (E); 2) German (G); 3) French (F); and 4) Japanese (J). For German, French, and Japanese reviews, the dataset provides their corresponding translated English reviews. Thus, we use English reviews as the source domain and the translated reviews from each language as the target domain in our experiments. For each category, there are 2000 positive and 2000 negative English reviews in the source domain, and 1000 positive and 1000 negative reviews for each of the other three languages in the target domain. All of the reviews are preprocessed using TF-IDF.

20 Newsgroup corpus is a widely used benchmark to evaluate transfer learning algorithms [1], [2], [4], [14], [16], [18], [39], which contain 20 000 documents distributed evenly in 20 different subcategories. We preprocess 20 Newsgroups by removing words occurring in less than eight documents. All of the documents are preprocessed using TF-IDF.

B. Baseline Algorithms

We compare our proposed method with the following state-of-the-art algorithms.

1) SVM and LR: Here, we adopt two supervised algorithms as the baseline methods, the SVM is trained with linear kernel because it achieves better performance on text classification tasks. Moreover, logistic regression is trained with default parameters.

2) Mtrick: It assumes the same association between the word features concepts and the example classes as the bridge across domains [4]. The number of word clusters k is set as 50.

3) DTL: The DTL method [18] simultaneously learns the marginal and conditional distributions, and it assumes that the word clusters can be partitioned into common and specific clusters. Like Mtric, the total number of word clusters k is set as 50, and the numbers of common and specific clusters are set as 25 and 25, respectively.

4) TriTL: The TRiTL [14], [16] method exploits both shared and distinct concepts for cross-domain text classification, which differs concepts into three groups: 1) identical concepts; 2) alike concepts; and 3) distinct concepts. As set in Mtrick and DTL, the number of word clusters k is equal to 50. As suggested by the authors, the number of identical concepts, alike concepts, and distinct concepts are set as 20, 20, and 10, respectively.

5) HIDC[2]*:* Concept learning for cross-domain text classification: a general probabilistic framework (HIDC) [23] method exploits the distinct concepts for cross-domain text classification. The number of word clusters k is set as 50, and the numbers of identical concepts, homogeneous concepts, and distinct concepts are set as 20, 20, and 10, respectively.

[1] http://www.uni-weimar.de/en/media/chairs/webis/corpora/

[2] http://www.intsci.ac.cn/users/zhuangfuzhen/

TABLE II
DETAILS OF THE 20 NEWSGROUP DATASETS

ID	source domain	subcategories	target domain	subcategories	# of training examples	# of test examples
1	comp	comp.graphics, comp.os rec.autos, rec.motorcycles	rec	comp.sys.ibm, comp.sys.mac rec.sport.baseball, rec.sport.hockey	3884	3901
2	comp	comp.graphics, comp.os sci.crypt,sci.med	sci	comp.sys.ibm, comp.sys.mac sci.electronics, sci.space	3883	3887
3	comp	comp.graphics, comp.os talk.politics.guns, talk.politics.mideast	talk	comp.sys.ibm, comp.sys.mac talk.politics.misc, talk.religion.misc	3753	3321
4	rec	rec.autos, rec.motorcycles sci.crypt, sci.med	sci	rec.sport.baseball, rec.sport.hockey sci.electronics, sci.space	3939	3930
5	rec	rec.autos, rec.motorcycles talk.politics.guns, talk.politics.mideast	talk	rec.sport.baseball, rec.sport.hockey talk.politics.misc, talk.religion.misc	3809	3364
6	sci	sci.crypt, sci.med talk.politics.guns, talk.politics.mideast	talk	sci.electronics, sci.space talk.politics.misc, talk.religion.misc	3808	3364

6) Cross-Lingual Structural Correspondence Learning[3]*:* Cross-lingual structural correspondence learning (CL-SCL) [37] first selects pivot features and then learns the structural correspondence information from the dataset. The number of pivot features is set as 50, which achieves the best performance on the datasets.

7) SCL-OM: This is an improved version of the CL-SCL method, which learns meaningful one-to-many mappings for pivot words by using large amounts of monolingual data and a small dictionary. Then, the authors proposed extending the pivots from one-to-one mapping in CL-SCL to one-to-many mappings in SCL-OM [40]. In our experiments, the total number of pivots is set as 600, which achieves the best performance on the cross-domain tasks.

8) Distribution Matching-Based Matrix Completion: The distribution matching-based matrix completion (DMMC) [38] method introduces a manual effort to acquire correspondence between heterogeneous domains. The parameters in DMMC are set as the authors suggested.

C. Implementation Details

The initializations of our approach and baselines are as follows.

1) F_{ss} and F_{sd} are initialized with the absolute values of normally distributed random numbers, respectively. Then, F_{ts} is initialized the same as F_{ss}, and F_{td} is initialized the same as F_{sd}. Finally, the four matrices are all column normalized. Note that these matrices of the other NMTF-based methods are initialized as the word clusters results by PLSA [41]. Specifically, $F_{s_{(ij)}}$ and $F_{t_{(ij)}}$ are both initialized as the output of PLSA on the whole dataset of the source and target domains. We adopt the MATLAB implementation of PLSA in the experiment.[4]
2) G_0 and G_s are initialized as the ground truth of the source label and M is initialized as $(G_s^\top G_s)^{-1} G_s^\top G_0$. G_t is initialized as the predicted results of any supervised classifier, which is trained based on the source domain data. In our experiment, logistic regression is adopted to give these initial results.
3) H is initialized as follows: each entry is assigned with the same value and the sum of values in each row satisfies $\sum_j H_{(ij)} = 1$.

Note that if the values of elements in matrix F for all of the NMF-based methods are less than 0 during the iteratively updating procedure, then they are set as 0. Mtric, DTL, TriTL, and our proposed method belong to transductive transfer learning, which do not use any additional data. CL-SCL, however, is a semisupervised transfer learning, because it uses a large amount of unlabeled data. DMMC is a kind of active learning, which needs manual efforts. The remaining parameter settings in the baseline algorithms and our method are as follows: the maximum iterative number of Mtric, DTL, TriTL, and our method in CDSC tasks are set as 180. After some preliminary tests, the parameters of our method are set as $\alpha = \beta = \gamma = \delta = 1.5$, $\pi = 0.01$, and $\epsilon = 10^{-5}$. We also provide the experiments of parameters analysis in the next section.

D. Experimental Design

For the Amazon reviews dataset, we conduct 18 kinds of cross-domain sentiment classification tasks.[5] Here, EB→FD uses English BOOKS reviews as the source domain and uses the translated DVD reviews from French as the target domain.

For 20 Newsgroup dataset, there are four top categories by combining similar subcategories, that is, *comp*, *sci*, *rec*, and *talk*. Similar to the previous works [4], any top two categories can be selected to construct the cross-domain dataset; thus, we conduct six groups experiments. The details of the constructed datasets are listed in Table II.

We use the classification accuracy on the test data (unlabeled target data) as the evaluation metric, since it is widely adopted in [4], [7], [18], [28], and [38]. Our method is the averaged accuracy of ten repeated experiments because of the random initialization of word clusters F.

VI. EXPERIMENTAL RESULTS

In this section, we will show the overall classification results, semantic analysis, and parameter study.

[3]https://github.com/pprett/nut

[4]http://www.kyb.tuebingen.mpg.de/bs/people/pgehler/code/index.html

[5]They are: EB→ FD, EB→ FM,ED→ FB, ED→ FM, EM→ FB, EM→ FD, EBGD, EB→ GM, ED→ GB, ED→ GM, EM→ GB, EM→ GD, EB→ JD, EB→ JM, ED→ JB, ED→ JM, EM→ JB, and EM→ JD.

TABLE III
CLASSIFICATION ACCURACY (%) ON AMAZON DATASETS (TEN REPEATED EXPERIMENTS FOR SA-TL)

Task	LR	SVM	Mtrick	TriTL	DTL	DMMC	HIDC	CL-SCL	SCL-OM	sa-TL
EB→ GD	77.43	75.43	79.36	76.98	75.62	78.28	77.92	79.17	79.12	**82.95***
EB→ GM	75.90	73.55	78.72	78.07	79.21	76.60	77.83	81.30	79.83	**83.96***
ED→ GB	78.80	76.30	80.19	81.03	80.87	77.47	79.10	**83.88**	82.09	83.25
ED→ GM	77.75	75.95	79.66	78.19	78.69	76.60	79.80	82.36	82.10	**84.74***
EM→ GB	76.75	74.15	78.72	79.43	77.68	77.47	78.18	**83.35**	83.06	83.02
EM→ GD	77.48	75.68	79.82	77.82	77.24	78.28	78.36	79.33	81.26	**81.36**
EB→ FD	79.95	77.30	82.44	77.97	80.69	76.23	80.78	76.84	79.27	**84.31***
EB→ FM	76.26	74.25	78.99	77.96	79.29	74.05	79.81	77.66	79.43	**84.14***
ED→ FB	76.41	74.76	79.86	80.23	79.85	76.52	81.57	80.93	81.85	**83.28***
ED→ FM	77.76	74.65	79.27	79.43	79.29	74.05	80.45	77.37	79.59	**83.09***
EM→ FB	76.66	74.56	79.91	77.85	77.20	76.52	80.23	80.13	80.81	**83.89***
EM→ FD	79.70	77.40	81.01	80.35	78.46	76.23	80.32	78.85	78.05	**83.98***
EB→ JD	71.39	70.11	73.03	70.41	70.11	72.12	73.53	75.78	74.70	**75.75**
EB→ JM	69.47	67.59	72.31	74.84	76.90	71.37	76.13	74.05	73.82	**81.27***
ED→ JB	69.87	68.37	71.71	69.61	70.99	68.54	72.62	74.92	73.95	**75.52**
ED→ JM	73.81	71.52	76.88	73.32	78.30	71.37	77.47	77.80	73.98	**81.41***
EM→ JB	68.37	67.13	69.91	68.13	67.21	68.54	69.55	69.41	**74.75**	73.03
EM→ JD	70.42	68.57	72.97	72.49	70.93	72.12	73.97	**76.24**	74.82	74.03
avg.	75.23	73.18	77.49	76.34	76.58	74.58	77.65	78.30	78.47	**81.31**

Note: *sa-TL performs significantly better than the second-best method using resampled paired t test.

TABLE IV
CLASSIFICATION ACCURACY (%) ON 20 NEWSGROUP DATASET (TEN REPEATED EXPERIMENTS FOR SA-TL)

source domain	target domain	LR	SVM	Mtrick	TriTL	DTL	HIDC	sa-TL
comp	rec	76.03	76.44	97.87	97.00	94.10	97.77	**98.75***
comp	sci	66.92	67.71	81.99	74.87	78.44	83.05	**87.84***
comp	talk	82.72	88.62	**98.49**	90.43	94.67	97.65	97.19
rec	sci	60.13	58.78	71.48	92.06	95.14	96.23	**97.27***
rec	talk	68.10	68.31	86.56	68.13	89.57	**97.18**	97.12
sci	talk	53.75	53.83	59.81	53.75	89.48	91.68	**95.56** *
avg.	-	67.94	68.95	82.70	79.37	90.23	93.93	**95.62**

Note: *sa-TL performs significantly better than the second-best method using resampled paired t test.

A. Classification Performance

In this section, we show the classification accuracies of 18 kinds of cross-domain sentiment classification tasks on Amazon datasets and six kinds of cross-domain text classification tasks on 20 Newsgroup dataset.

Table III shows the average classification accuracies on 18 kinds of cross-domain sentiment classification tasks.

1) Generally, as shown in Table III, we can observe that our proposed method performs consistently best in terms of averaged accuracy. Our method obtains the best accuracy in 15 out of 18 tasks. The last row of Table III is the average accuracy, and we observe that our proposed method achieves the best performance in general.
2) The semisupervised SCL-OM, CL-SCL, Mtrick, and HIDC achieve better performance than the other methods, but they are inferior to our method. SVM, DMMC, and LR obtain the worst accuracy, and we can find that traditional classification methods (LR and SVM) perform worse than the transductive transfer learning methods.
3) DTL and TriTL acquire the second worst results, because both of them assume that the cluster association matrix is kept unchanged in the cross-domain tasks. In fact, the results also demonstrate that the assumption is not satisfied. As shown in Table III, for transductive transfer learning, our method improves up to 5.0%, 4.7%, 3.8%, and 3.7%, respectively, compared with TriTL, DTL, Mtrick, and HIDC.
4) The improvement is also up to 7% compared with the active learning method DMMC, which uses additional unlabeled data to learn correspondences. Our proposed method also achieves better performance than the semisupervised CL-SCL method and its variant named SCL-OM, which both need an additional 50 000 unlabeled reviews. Our method, however, only applies to the source domain training data. We also observe that the SCL-OM method performs slightly better than CL-SCL because the former extends pivots by word2vec and learns more of a semantic relationship between the source domain and target domain. In addition, we also conduct a series of ten trial experiments to verify the significance between the second-best method and our proposed sa-TL. In the resampled paired t test [42], our proposed ssSCL-ST outperforms the second-best method significantly (p-value < 0.01) on 11 out of 18 kinds of cross-domain sentiment classification tasks for the Amazon reviews dataset.

For 20 Newsgroup, DMMC, CL-SCL, and SCL-OM methods are excluded because this dataset provides no auxiliary data. Here, we show the classification accuracies of the transductive transfer learning methods on six kinds of cross-domain text classification tasks in Table IV.

1) We can observe that our proposed method achieves the best averaged accuracy for the 20 Newsgroup dataset. Our sa-TL method performs significantly better than the remaining methods, except for HIDC.

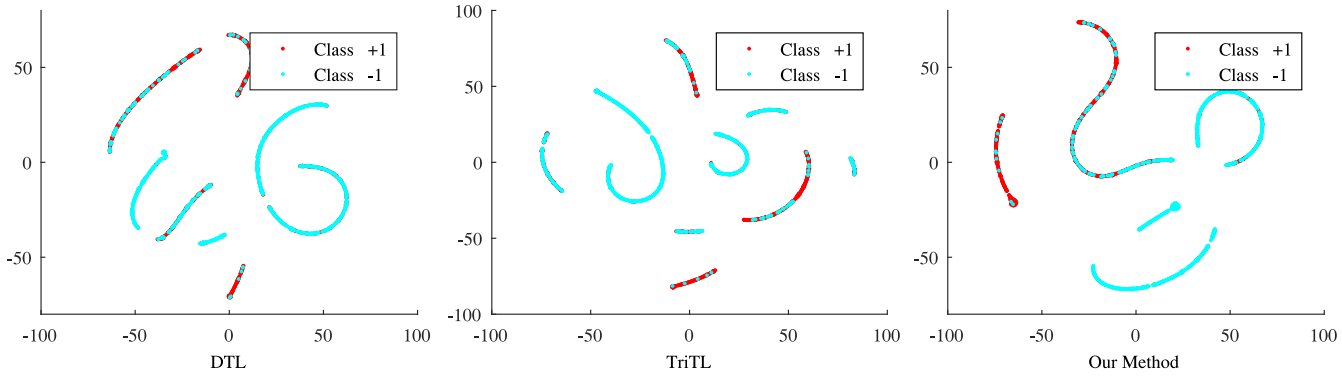


Fig. 1. 2-D embedding visualization of topics on different classes (best viewed in color).

TABLE V
TOP WORDS OF SIMILAR AND DISTINCT CONCEPTS ON THE EB→ GD TASK

Concepts ID	Domain	Top words
Topic11 (similar concept)	English Book	fantastic best favorite touching great recommend classic fast year life must shipping makes wonderful child mind
Topic11 (similar concept)	Germany DVD	fantastic best great favorite touching shipping recommend must year fast classic life makes wonderful who child
Topic23 (distinct concept)	English Book	bore disappointing money disneyland pages impossible poor sounded unbelievable folks stinks information offensive author magazine
Topic48 (distinct concept)	Germany DVD	disappointing yawn concoction poorly cheap trash bad unfortunately deserves bored films implementation revolutions weak wooden

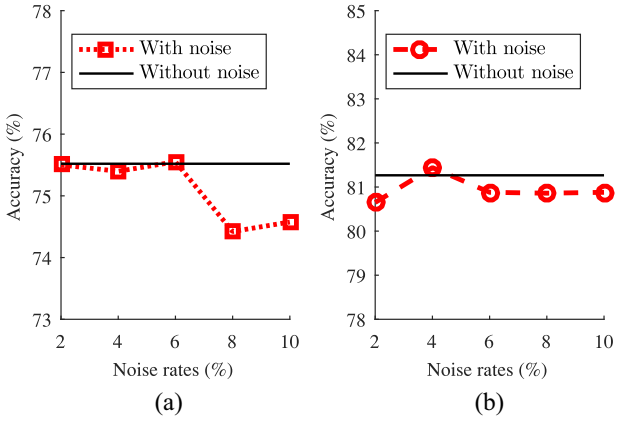


Fig. 2. Accuracy with different noise rates. (a) EBJD. (b) EBJM.

2) LR and SVM perform worst on 20 Newsgroup, because they completely ignore the knowledge transfer across different domains.
3) For transductive transfer learning methods, the accuracies of DTL, HIDC, and our method are much higher than those of the Mtric and TriTL methods. We also conduct a series of ten trial experiments to verify the significance between the second-best method and sa-TL. In the resampled paired t test [42], we can observe that sa-TL outperforms the second-best method significantly (p-value < 0.01) on 4 out of 6 tasks for the 20 Newsgroup dataset.

From different classification tasks, we can observe that our method performs better than the state-of-the-art methods by allowing differences of word clusters matrices and clusters association matrices. It shows that sa-TL is more flexible when the source and target domain examples own their different feature distributions.

B. Topic Analysis

We demonstrate the virtues of mined topics by our proposed sa-TL method from two aspects, that is, the semantic representation of topics and the visualized embedding of topics. The former shows that the mined topics are meaningful, and the latter illustrates that the mined topics help to form a better grouping and separation for classification.

Table V shows a sample of similar and distinct concepts between the source domain and the target domain on the EB→ GD dataset. For similar concepts, we can observe that most of the top words are the same, but the weight of each word in similar topics is different. These similar concepts are the knowledge bridge between the source domain and the target domain. Moreover, we also select one distinct concept from each domain. For the source domain, we can find that the negative concept "Topic23" is about a book, whereas the negative concept "Topic48" talks more about film. Through the sample, we can observe that our approach not only learns better similar concepts between two domains but also detects the domain-specific concepts.

Rather than the same word clusters matrices and clusters association matrices, we argue that the softly associative regularization constraints should be added to allow differences while keeping the knowledge transferring function. For cross-domain text classification tasks, we can observe the difference between our method and DTL/TriTL by presenting the visualization analysis of topics after different matrix factorization, which will allow us to understand how crucial the assumption of the approximation between two association matrices is. In NMTF, we can use HG as the approximately inferred topic proportions because F is the word-topic distribution. Therefore, we present an empirical assessment of topic estimation on the target domain MUSIC reviews, and we compare the relations between topic distributions and class labels. The number of clusters is set as 50.

Fig. 1 (best viewed in color) shows the 2-D embedding of the approximately inferred topic proportions (HG) by our

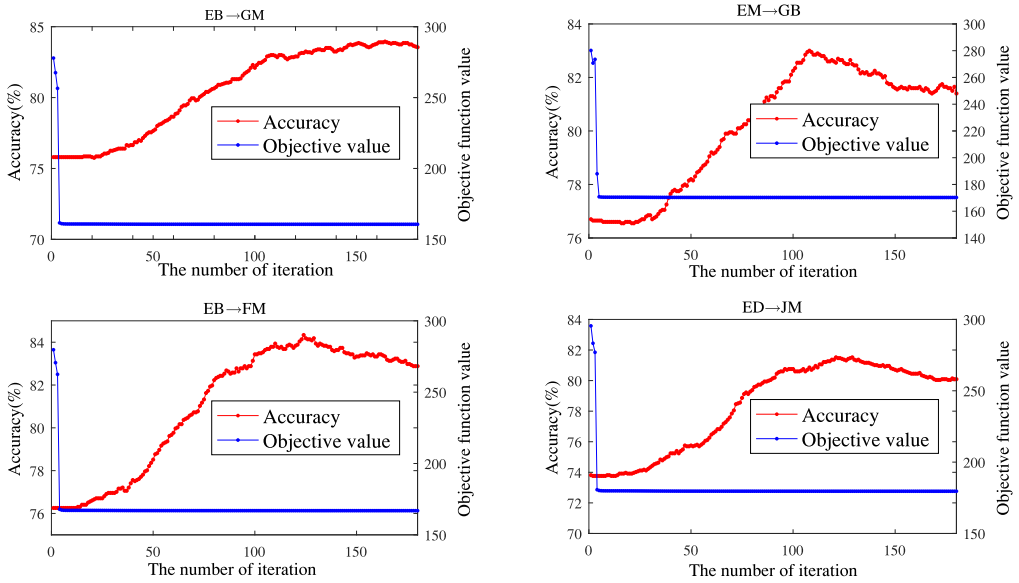


Fig. 3. Convergence verification.

method, DTL, and TriTL, respectively. The t-SNE stochastic neighborhood embedding method [43] is shown in Fig. 1. Each dot represents a document and each color-shape pair represents a category. Visually, our proposed method is apt to produce a better grouping and separation of the documents in positive and negative categories (represented by the red dot and green dot). In contrast, DTL and TriTL do not produce a well-separated embedding, and documents in different categories tend to mix together. We can observe that the mixed embedding in DTL and TriTL is similar, which leads to similar classification accuracy. Note that there exist some documents with positive or negative labels which are also mixed together in Fig. 1, and the mixture is caused by the closer semantic distance among documents and the low-dimensional projection. Intuitively, a well-separated representation is more discriminative for text categorization. Therefore, compared with traditional transfer learning methods, our proposed sa-TL model can achieve better classification accuracy.

C. Noisy Label

In our sa-TL method, we introduce the M matrix to learn the best mapping between G_0 and G_s, and the term $\pi||G_0-G_sM||^2$ works as a regularizer to measure the similarity between the ground truth of source label G_0 and the learned G_s in the partition level, by which our method can adapt the noisy label in the source domain. Here, the noisy label refers to the sample labels, which are randomly corrupted [44] and may come from the hand-labeled data or the noise of data. In our experiments, we select two tasks (EB→ JD and EB→ JM) from the Amazon dataset and reverse the class label of examples randomly as the noisy label with different noise rates (from 2% to 10%). As shown in Fig. 2, the black curve is the accuracy of sa-TL when the training data do not contain any noise labels, which is also the comparative baseline, and the dotted red curve is the accuracy achieved by our proposed sa-TL with different noise rates. We can observe that the performance of our sa-TL decreases slightly when the noise rate is bigger than 8%, and the difference is only 1%. It demonstrates that our proposed sa-TL has better robustness for CDC tasks by introducing the mapping matrix M.

D. Convergence Verification

We also empirically verify the convergence of our proposed iteration algorithm, as shown in Fig. 3. Since the word clusters matrices F_s and F_t are both random initializations, which will disturb the value of the objective function at the first iteration, we thus omit the values in Fig. 3. We can see that the value of the objective function (the right y-axis) decreases quickly along with iterative steps for the randomly chosen tasks. The results are consistent with our theoretic analysis. The left y-axis in Fig. 3 is the classification accuracy, and we can observe that the maximum iteration number changes with different datasets. For example, for EM→ GB and EB→ FM tasks, the iteration number should not exceed 120; otherwise, it will be overfitting.

E. Parameter Study

At last, we study the influences of parameters in our proposed method. They are: α, β, γ, δ, π, the total number of topics, and the number of similar topics k_1. Generally, the value of k is data-dependent; thus, we focus on the remaining five parameters. We randomly sample the values of α, β, γ, δ in (0.5, 10), and the value of k_1 in (10, 50]. The parameter π is chosen among six values between 1E-5 and 1. Table VI shows the accuracy influence of the parameters on randomly chosen six kinds of cross-domain sentiment classification tasks. The 16 groups of sampling experiments demonstrate that our method is not sensitive to the parameters setting when their bounds are predefined properly, because the mean is very close to that using the default parameters and the variance is very small. Therefore, we set α, β, γ, and δ as the same value in this paper.

TABLE VI
PARAMETRIC INFLUENCE TO THE ACCURACY (%) OF SA-TL

sample ID	α	β	γ	δ	π	k_1	EB→ GM	ED→ GM	EB→ FD	EM→ FD	EB→ JM	ED→ JM
1	6.30	0.87	3.76	9.24	1E-03	40	84.25	84.70	84.45	84.00	81.57	81.62
2	9.96	0.75	4.97	1.48	1E-04	33	84.20	84.45	84.50	84.00	81.42	81.01
3	2.19	3.46	9.30	9.75	1E-05	31	84.25	84.55	84.25	84.05	81.27	81.57
4	6.55	7.01	8.20	1.91	1E-03	38	83.70	84.65	84.55	84.15	81.37	81.32
5	2.40	2.15	1.49	0.99	1E-04	17	83.80	84.60	84.50	84.15	81.01	81.06
6	3.89	3.97	2.29	4.12	1E-05	21	84.05	84.65	84.30	83.95	81.57	81.37
7	1.22	1.60	7.86	6.05	1E-03	43	84.05	84.50	84.35	83.90	81.27	81.42
8	0.63	3.72	2.97	4.65	1E-05	40	84.00	84.60	84.40	83.85	81.11	81.42
9	3.15	5.70	9.60	9.67	1	17	83.60	84.65	84.30	83.90	81.62	81.52
10	8.35	8.72	2.66	4.61	0.1	20	83.70	84.60	84.15	84.15	81.37	81.47
11	2.53	8.09	8.73	8.54	1E-03	19	84.00	84.50	84.70	83.95	81.01	81.27
12	7.51	6.34	1.83	4.93	1	21	83.75	84.75	84.65	84.10	81.62	81.57
13	6.10	0.99	5.07	9.76	0.1	21	84.05	84.60	84.35	84.00	81.83	81.47
14	2.71	2.83	6.77	9.24	1E-05	36	84.25	84.70	84.35	84.05	81.42	81.47
15	0.86	5.78	1.67	8.52	1E-04	36	83.95	84.75	84.35	84.20	81.57	81.52
16	5.99	2.06	5.88	9.79	1	46	84.10	84.65	84.50	84.10	81.57	81.42
Mean	-	-	-	-	-	-	83.98	84.62	84.42	84.03	81.41	81.41
Variance	-	-	-	-	-	-	0.207	0.085	0.141	0.101	0.227	0.167
This Paper	1.5	1.5	1.5	1.5	0.01	20	83.92	84.68	84.37	83.96	81.46	81.63

VII. CONCLUSION

In this paper, we propose a novel transductive transfer learning algorithm by introducing softly associative regularizations on the word clusters matrix and clusters association matrix and allowing label noise in the source domain for CDC. We presented a joint optimization framework of two matrix tri-factorizations, and derive an efficient iterative algorithm with its convergence verified empirically. We conducted a set of binary classification tasks to evaluate the effectiveness of our proposed algorithm. The experimental results showed that our approach outperforms abundant state-of-the-art methods. In the future, we will consider applying our model to multiclasses CDC tasks.

REFERENCES

[1] J. Gao, W. Fan, J. Jiang, and J. Han, "Knowledge transfer via multiple model local structure mapping," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min. (KDD)*, 2008, pp. 283–291.

[2] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[3] J. Pan, G.-R. Xue, Y. Yu, and Y. Wang, "Cross-lingual sentiment classification via bi-view non-negative matrix tri-factorization," in *Proc. Adv. Knowl. Disc. Data Min.*, 2011, pp. 289–300.

[4] F. Zhuang *et al.*, "Exploiting associations between word clusters and document classes for cross-domain text categorization," in *Proc. SDM*, 2010, pp. 13–24.

[5] H. Wang, H. Huang, F. Nie, and C. Ding, "Cross-language Web page classification via dual knowledge transfer using nonnegative matrix tri-factorization," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retriev.*, 2011, pp. 933–942.

[6] X. Meng *et al.*, "Cross-lingual mixture model for sentiment classification," in *Proc. Annu. Meeting Assoc. Comput. Linguist. (ACL)*, 2012, pp. 572–581.

[7] G. Zhou, T. He, J. Zhao, and W. Wu, "A subspace learning framework for cross-lingual sentiment classification with partial parallel data," in *Proc. Int. Joint Conf. Artif. Intell.*, Buenos Aires, Argentina, 2015, pp. 1426–1432.

[8] J. Li, Y. Wu, J. Zhao, and K. Lu, "Low-rank discriminant embedding for multiview learning," *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3516–3529, Nov. 2017.

[9] M. Jiang, W. Huang, Z. Huang, and G. G. Yen, "Integration of global and local metrics for domain adaptation learning via dimensionality reduction," *IEEE Trans. Cybern.*, vol. 47, no. 1, pp. 38–51, Jan. 2017.

[10] Y. Lin *et al.*, "Cross-domain recognition by identifying joint subspaces of source domain and target domain," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 1090–1101, Apr. 2017.

[11] J. Li, K. Lu, Z. Huang, L. Zhu, and H. T. Shen, "Transfer independently together: A generalized framework for domain adaptation," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2018.2820174.

[12] J. Li, K. Lu, Z. Huang, L. Zhu, and H. T. Shen, "Heterogeneous domain adaptation through progressive alignment," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2018.2868854.

[13] Z. Ding, M. Shao, and Y. Fu, "Incomplete multisource transfer learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 2, pp. 310–323, Feb. 2018.

[14] F. Zhuang *et al.*, "Triplex transfer learning: Exploiting both shared and distinct concepts for text classification," *IEEE Trans. Cybern.*, vol. 44, no. 7, pp. 1191–1203, Jul. 2014.

[15] W. Dai, G.-R. Xue, Q. Yang, and Y. Yu, "Co-clustering based classification for out-of-domain documents," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2007, pp. 210–219.

[16] F. Zhuang, P. Luo, C. Du, Q. He, and Z. Shi, "Triplex transfer learning: Exploiting both shared and distinct concepts for text classification," in *Proc. WSDM* , 2013, pp. 425–434.

[17] S. Li, Y. Xue, Z. Wang, and G. Zhou, "Active learning for cross-domain sentiment classification," in *Proc. 23rd Int. Joint Conf. Artif. Intell. (IJCAI)*, 2013, pp. 2127–2133.

[18] M. Long *et al.*, "Dual transfer learning," in *Proc. SDM*, 2012, pp. 1–12.

[19] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, "Cross-domain sentiment classification via spectral feature alignment," in *Proc. 19th Int. Conf. World Wide Web (WWW)*, 2010, pp. 751–760.

[20] L. Gui *et al.*, "Cross-lingual opinion analysis via negative transfer detection," in *Proc. Annu. Meeting Assoc. Comput. Linguist. (ACL)*, Jun. 2014, pp. 860–865.

[21] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification," in *Proc. 45th Annu. Meeting Assoc. Comput. Linguist. (ACL)*, Jun. 2007, pp. 440–447.

[22] T. Li, V. Sindhwani, C. Ding, and Y. Zhang, "Knowledge transformation for cross-domain sentiment classification," in *Proc. SIGIR*, 2009, pp. 716–717.

[23] F. Zhuang, P. Luo, P. Yin, Q. He, and Z. Shi, "Concept learning for cross-domain text classification: A general probabilistic framework," in *Proc. IJCAI*, 2013, pp. 1960–1966.

[24] H. Wang and Q. Yang, "Transfer learning by structural analogy," in *Proc. 25th AAAI Conf. Artif. Intell. (AAAI)*, 2011, pp. 513–518.

[25] X. Hu *et al.*, "Multi-bridge transfer learning," *Knowl. Based Syst.*, vol. 97, pp. 60–74, Apr. 2016.

[26] B. Tan, Y. Song, E. Zhong, and Q. Yang, "Transitive transfer learning," in *Proc. ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2015, pp. 1155–1164.

[27] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2004, pp. 109–117.

[28] M. Long, J. Wang, G. Ding, D. Shen, and Q. Yang, "Transfer learning with graph co-regularization," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 7, pp. 1805–1818, Jul. 2014.

[29] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.
[30] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 556–562.
[31] J. Yoo and S. Choi, "Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on Stiefel manifolds," *Inf. Process. Manag.*, vol. 46, no. 5, pp. 559–570, 2010.
[32] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix *t*-factorizations for clustering," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2006, pp. 126–135.
[33] H. Liu, M. Shao, and Y. Fu, "Structure-preserved multi-source domain adaptation," in *Proc. IEEE 16th Int. Conf. Data Min. (ICDM)*, Dec. 2016, pp. 1059–1064.
[34] B. Mirkin, "Reinterpreting the category utility function," *Mach. Learn.*, vol. 45, no. 2, pp. 219–228, 2001.
[35] J. Wu, H. Liu, H. Xiong, J. Cao, and J. Chen, "K-means-based consensus clustering: A unified view," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 1, pp. 155–169, Jan. 2015.
[36] P. Prettenhofer and B. Stein, "Cross-language text classification using structural correspondence learning," in *Proc. Annu. Meeting Assoc. Comput. Linguist. (ACL)*, 2010, pp. 1118–1127.
[37] P. Prettenhofer and B. Stein, "Cross-lingual adaptation using structural correspondence learning," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 1, pp. 1–22, 2011.
[38] J. T. Zhou, S. J. Pan, I. W. Tsang, and S.-S. Ho, "Transfer learning for cross-language text categorization through active correspondences construction," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2016, pp. 2400–2406.
[39] D. Zhang, J. He, Y. Liu, L. Si, and R. Lawrence, "Multi-view transfer learning with a large margin approach," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min. (KDD)*, 2011, pp. 1208–1216.
[40] N. Li, S. Zhai, Z. Zhang, and B. Liu, "Structural correspondence learning for cross-lingual sentiment classification with one-to-many mappings," in *Proc. 31st Conf. Artif. Intell. (AAAI)*, San Francisco, CA, USA, Feb. 2017, pp. 3490–3496.
[41] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, nos. 1–2, pp. 177–196, Jan. 2001.
[42] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Dec. 2006.
[43] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[44] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 447–461, Mar. 2016.

Deqing Wang (M'19) received the Doctoral degree in computer science from Beihang University, Beijing, China, in 2013.

He is an Assistant Professor with the School of Computer Science, Beihang University, where he is the Deputy Chief Engineer with the National Engineering Research Center for Science Technology Resources Sharing and Service. His current research interests include text categorization and data mining for software engineering and machine learning.

Chenwei Lu is currently pursuing the master's degree in computer science with Beihang University, Beijing, China.

His current research interests include transfer learning and sentiment analysis.

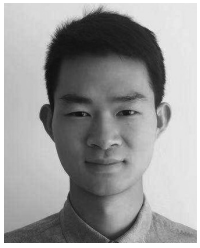
Junjie Wu received the Ph.D. degree in management science and engineering from Tsinghua University, Beijing, China, in 2008.

He is currently a Full Professor with the Information Systems Department, School of Economics and Management, Beihang University, Beijing, where he is also the Director of Research Center for Data Intelligence and the Vice Director of the Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations. His current research interests include data mining, with a special interest in social, urban, and financial computing.

Mr. Wu was a recipient of the various national academic awards, including the NSFC Distinguished Young Scholars, the MOE Changjiang Young Scholars, and the National Excellent Doctoral Dissertation.

Hongfu Liu received the bachelor's and master's degrees in management information systems from the School of Economics and Management, Beihang University, Beijing, China, in 2011 and 2014, respectively, and the Ph.D. degree in computer science from Northeastern University, Boston, MA, USA, in 2018.

He is currently a Tenure-Track Assistant Professor of computer science with Brandeis University, Waltham, MA, USA. His current research interests include data mining and machine learning, with special interests in ensemble learning.



Wenjie Zhang received the master's degree in computer science from Beihang University, Beijing, China, in 2016.

He is currently an Engineer with Yidian News Inc., Beijing. His current research interests include transfer learning and information recommendation.

Fuzhen Zhuang received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2011.

He is currently an Associate Professor with the Institute of Computing Technology, Chinese Academy of Sciences. He has published over 80 papers in some prestigious refereed journals and conference proceedings, such as the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the IEEE TRANSACTIONS ON CYBERNETICS, *ACM Transactions on Intelligent Systems and Technology*, and *Information Sciences*. His current research interests include transfer learning, machine learning, data mining, multitask learning, and recommendation systems.

Hui Zhang received the M.S. and Ph.D. degrees in computer science from Beihang University, Beijing, China, in 1994 and 2009, respectively.

He is currently a Professor with the State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University. He is currently with a Chinese Science and Technology Resources Portal, Beijing, as a Chief Architect. His current research interests include cloud computing, web information retrieval, and data mining.