Bayesian Bridging Topic Models for Classification

MENG-SUNG WU
Computational Intelligence Technology Center
Industrial Technology Research Institute
Hsinchu, 310 Taiwan

We study the problem of constructing the topic-based model over different domains for text classification. In real-world applications, there are abundant unlabeled documents but sparse labeled documents. It is challenging to construct a reliable and adaptive model to classify a large amount of documents containing different domains. The classifiers trained from a source domain shall perform poorly for the test data in a target domain. Also, the trained model is vulnerable to the weakness of classification among ambiguous classes. In this study, we tackle the issues of domain mismatch and confusing classes and conduct the discriminative transfer learning for text classification. We propose a Bayesian bridging topic models (BTM) from a variety of labeled and unlabeled documents and perform the transfer learning for cross-domain text classification. A structural model is built and its parameters are estimated by maximizing the joint marginal likelihood of labeled and unlabeled data via a variational inference procedure. We also construct the discriminative learning on our proposed model for adjust parameters by using the minimum classification error criterion. We show that improvements over cross-domain text classification using the proposed model can be achieved better performance than other models.

Keywords: transfer learning, topic model, cross-domain classification, latent Dirichlet allocation, Bayesian

1. INTRODUCTION

Supervised learning plays a crucial role in pattern recognition and machine learning. Frameworks for supervised learning usually require a large number of labeled examples to identify classification models with an acceptable degree of reliability [19]. However, labeled data is often unavailable due to the prohibitive cost associated with the labeling process. In addition, classifiers trained in one domain are not necessary applicable in other domains. Thus, to avoid mismatch between training sets and test data distribution, classifiers should be designed to accommodate a range of domains. Transfer learning [8, 10] has previously been proposed to overcome domain mismatch and insufficiencies in labeled data. Transfer learning refers to applying knowledge gained from one task to a different but related task. Adaptive learning is used to transfer knowledge from labeled source data to unlabeled target data in order to improve the generalizability of learning tasks in cross-domain classification. For example, learning to recognize categories within the field of computers might help a classifier recognize categories in other fields. Given labeled data from a source domain and unlabeled data from a target domain, it may be possible to predict labels for the target domain by employing labels and data from the source domain. Ando and Suzuki [1] and Dai et al. [7] employed a co-clustering algorithm [9] in the transfer of class labels and knowledge from a source domain to a
target domain for cross-domain text classification. Likewise, Dai et al. [8] modified the native Bayes algorithm to address domain adaptation by employing the expectation-maximization (EM) algorithm to form a locally optimal posterior hypothesis according to test data distribution. Do and Ng [10] proposed a meta-learn algorithm to estimate the model parameters from a set of related classification tasks. However, none of these studies involved transfer learning in a latent feature space. Xue et al. [29] addressed cross-domain text classification that used the topics to bridge the domain transfers by extending the traditional probabilistic topic model. In addition, Grandvalet and Bengio [11] proposed semi-supervised logistic regression by incorporating unlabeled data in the training of logistic regressions. Joachims [14] presented transductive inference for support vector machines (SVM) that incorporated knowledge from specific test data into SVM optimization. Nigam et al. [21] combined the EM algorithm and naïve-Bayes classifier for learning from both labeled and unlabeled documents under a semi-supervised learning framework.

Numerous studies have recently been proposed and have showed that discriminative learning has been successful in various applications of machine learning. Lacoste-Julien et al. [16] proposed a discriminative method for classification based on latent Dirichlet allocation (LDA) [3] where Gibbs sampling was applied for training LDA. This method was used to learn a projection matrix to transform topic assignments into labeled assignments. Zhu et al. [31] presented the maximum entropy discriminative LDA, which integrated the max-margin principle, to train the supervised topic model. Bickel et al. [2] focused on the mismatch between the training and test distribution by integrating weight estimation and model training in a unified discriminative framework. In addition, various successful discriminative techniques have been applied to speech recognition [13], such as the training criterion for maximum mutual information (MMI) and minimum classification error (MCE). Furthermore, a popular discriminative model was developed for an information retrieval and language model applying both the maximum entropy model [20] and minimum rank error model [5], individually.

For this study, we developed a Bayesian topic-based model for transfer learning. The latent topic model is useful for compact modeling and for bridging different domains for knowledge transfer. In the literature, probabilistic latent semantic analysis (PLSA) [12] and LDA [3] are two latent topic paradigms that have been extensively applied in speech recognition [6, 23] and information retrieval [4, 26]. In PLSA, each document is characterized by a mixture model containing latent semantic mixtures. The parameters of the mixture probabilities were estimated by the maximum likelihood (ML) principle. In LDA, each document is viewed as a mixture distribution over latent topics. However, these models were designed to build a document model from a single data collection; thus, distributions of different domain data could not be assumed differently. Therefore, these methods were not directly adopted for transfer learning. We focus on adaptive learning across different domains consisting of labeled (source) and unlabeled (target) documents. We develop a Bayesian bridged topic model (BTM) for transfer learning. Statistical cross-domain learning is performed for document classification based on two related domains that could be generated by the shared topics. Xue et al. [29] presented the topic-bridged PLSA (TPLSA) model to bridge from the source domain to the target domain. Certain additional constraints were imposed for clustering by using prior knowledge from the training data. The proposed BTM model is more general than the
BAYESIAN BRIDGING TOPIC MODELS FOR CLASSIFICATION

TPLSA model because the BTM has the capability of predicting new documents. We use a variational Bayesian procedure to locate the empirical posterior distribution of latent variables, which approximates the true posterior distribution and obtains the BTM parameters. We further improve the performance of document classification by discriminative learning for BTM (DBTM) where the cross-domain classification of ambiguous classes is addressed using the discriminative model rather than the generation model. This DBTM is trained according to the MCE objective function, which is expressed as a continuous-valued classification error, or the expected loss function calculated using the discriminative functions of the target class model relative to the competing class models. We built the DBTM by minimizing the expected loss function through a generalized probabilistic descent algorithm [13]. The optimal Bayes decision is fulfilled for discriminative transfer learning. In the experiments, we implement the DBTM and demonstrate its superiority for cross-domain document classification by using the 20 Newsgroup dataset.

First, we present the principles of LDA and describe how TPLSA is applied to transfer learning. Second, we present the BTM with the variational inference algorithm and introduce the DBTM using the discriminative learning approach. Finally, we offer a conclusion and discuss future research directions.

2. SURVEY OF RELATED WORKS

2.1 Latent Dirichlet Allocation

Blei et al. [3] exploited the latent Dirichlet allocation (LDA) method to deal with the problem of PLSA that the aspect models were estimated from the observed training documents. LDA is a paradigm of document model where the topics are characterized by a distribution over words, and the documents are sampled from a mixture of topics. Each word \( w \) in a document \( d \) is associated with a latent topic \( z \). For each document, a multinomial distribution \( \theta \) over topic \( z \) is sampled from a Dirichlet distribution with parameter \( \alpha \). The topic is chosen by the topic mixture \( \theta \). A word is sampled from the multinomial distribution given a topic \( z \) and a parameter \( \beta \). The joint probability of \( w = \{w_n\}, z = \{z_n\} \) and \( \theta \) is expressed by

\[
p(w, z, \theta \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{d=1}^{N_d} \prod_{n=1}^{N_n} p(w_n \mid z_n, \beta) p(z_n \mid \theta). \tag{1}
\]

The marginal distribution of a document is represented by summing out the hidden variable \( z \) and integrating over \( \theta \), as follows

\[
p(w \mid \alpha, \beta) = \int p(\theta \mid \alpha) \prod_{d=1}^{N_d} \prod_{n=1}^{N_n} p(w_n \mid z_n, \beta) p(z_n \mid \theta) d\theta.
\tag{2}
\]

The marginal likelihood \( p(w \mid \alpha, \beta) \) over latent variables \( z \) and \( \theta \) is then calculated and maximized to find the LDA parameters \( \alpha \) and \( \beta \). However, the direct optimization is intractable. The variational inference [3, 30] and Gibbs sampling [27] are useful for model
inference. More specifically, we can adopt a variational Bayesian EM (VB-EM) algorithm [3] by maximizing the lower bound of marginal likelihood of Eq. (1). Variational distribution was used for latent variables \( (z, \theta) \)

\[
q(z, \theta \mid \eta, \phi) = q(\theta \mid \eta) \prod_{n=1}^{N} q(z_n \mid \phi_n),
\]

where \( \eta \) and \( \phi \) are variational Dirichlet parameter and multinomial parameter, respectively. In VB-E step, we estimate the variational parameters \( \phi \) and \( \eta \) corresponding to latent variables \( z \) and \( \theta \), respectively. By applying \( \phi \) and \( \eta \), LDA parameters \( \alpha \) and \( \beta \) are estimated in VB-M step. The VB-EM steps are assured to achieve the maximum of lower bound. Hence, the variational parameters were determined by

\[
\phi_{nk} \propto \beta_{nk} \exp \{\Psi(\eta_k) - \Psi(\sum_{j=1}^{K} \eta_j)\}
\]

\[
\eta_k = \alpha_k + \sum_{n} \phi_{nk}
\]

where \( \Psi(.) \) is the digamma function. Eqs. (4)-(5) are applied iteratively until convergence condition is met. In M-step, the conditional multinomial parameters are updated by

\[
\beta_{ij} \propto \sum_{m} \sum_{n} \phi_{n,m}^{ij} w_{mi}.
\]

Parameters of Dirichlet prior \( \alpha \) are estimated by the Newton-Raphson or gradient descent procedure.

2.2 Topic-Bridged PLSA

To establish a model across different data collections, the topic-bridged PLSA (TPLSA) [29] was proposed. This model was extended from PLSA [12] by integrating labeled and unlabeled data. The hidden variable \( z \) was used to bridge two domains. A cross-domain text classification was realized by adopting the shared probability of topic given word \( p(z \mid w) \) in calculation of document probability

\[
p(d \mid w) = \sum_{z} p(d \mid z) p(z \mid w)
\]

where \( d \in \{d_l, d_u\} \). The domains of source documents \( D_s = \{d_l\} \) and target documents \( D_t = \{d_u\} \) were assumed to be relevant in the probabilities \( p(d_l \mid z) \) and \( p(d_u \mid z) \) respectively. The source and target documents were merged into an integrated model. The objective function of TPLSA was established by considering the must-link and cannot-link constraints [24] that two source documents \( \{d_l^i, d_l^j\} \) were on the same topic \( z \) and on different topics \( \{z, z_i\} \). The expectation of likelihood function was calculated as

\[
\sum_{w} \lambda \sum_{d_l} n(w, d_l) \log \sum_{z} p(d_l \mid z) p(z \mid w) + (1 - \lambda) \sum_{w} n(w, d_u)
\]
\[
\log(\sum_{z} p(d_i | z) p(z | w)) + \nu_i \log(\sum_{z} p(d'_i | z) p(d'_i | z))
\]

where \(n(w, d)\) is the number of times \(w\) appears in the document \(d\), \(\lambda\) is a weight between labeled and unlabeled data, and \(\{\nu_1, \nu_2\}\) denote the weights for the constraints during maximum likelihood (ML) estimation. Due to the latent variable \(z\) in TPLSA, EM algorithm should be applied to estimate the ML parameters. The topic distribution was shared across different domains, over a fixed set of documents. The new TPLSA parameters are obtained by \[29\]

\[
p(d_i | z) \propto \sum_{w} n(w, d_i) p(z | d_i, w)
\]

\[
p(d'_i | z) \propto \sum_{w} n(w, d'_i) p(z | d'_i, w) + \nu_i \sum_{d_i \in D_i} p(z | d_i, d'_i) + \nu_i \sum_{d_i \in D_i, z \neq z'} p(z, z' | d_i, d'_i)
\]

\[
p(z | w) \propto \lambda \sum_{d_i} n(w, d_i) p(z | d_i, w) + (1 - \lambda) \sum_{d_i} n(w, d_i) p(z | d_i, w)
\]

where the posterior probability using current estimate \(\theta\) is expressed by

\[
p(z | d_i, w) = \frac{p(z | w) p(d_i | z)}{\sum_z p(z | w) p(d_i | z)}
\]

\[
p(z | d_i, d'_i) = \frac{p(d_i | z) p(d'_i | z)}{\sum_z p(d_i | z) p(d'_i | z)}
\]

\[
p(z, z' | d_i, d'_i) = \frac{p(d_i | z) p(d'_i | z')}{\sum_{z, z'} p(d_i | z) p(d'_i | z')}
\]

### 3. BAYESIAN BRIDING TOPIC MODEL

#### 3.1 Model Construction

To tackle the cross-domain modeling via transfer learning, we extend the LDA model, which was trained from single data collection, to a bridged topic model (BTM), which is jointly estimated from source data \(D_s = \{d_i\}\) and target data \(D_t = \{d'_i\}\), for transfer learning. The bridged topic model is depicted in Fig. 1. Each source document \(d_i\) with \(N'\) words is labeled by a class \(c\) with probability \(\theta \sim p(\theta | \alpha_c, c)\) while each target document \(d'_i\) with \(N'_\) words has no class label. Each word \(w_e\) in \(d \in \{d_i, d'_i\}\) is drawn from topic-dependent word distributions \(p(w_e = v | z_k = k, \beta)\) with a parameter \(\beta = \{\beta_k\}\), shared for topics in source and target domains. The topics \(\{z_k\}\) in both domains were...
assigned using different topic mixtures $\theta_l$ and $\theta_u$ which are Dirichlet distributed with corpus-level parameters $\alpha_l$ and $\alpha_u$, respectively. The topic mixture $\theta_l$ is class-dependent with a set of $C$ Dirichlet priors $\theta_l \sim p(\theta_l | \alpha_l, c_l)$ while $\theta_u$ is drawn from a single Dirichlet prior $\theta_u \sim p(\theta_u | \alpha_u)$. The topic $z^l_w$ of a target word $w^l_u$ is driven by $\theta_l$ using $p(z^l_w | \theta_l)$. However, the topic $z^u_w$ of word $w^u_u$ in source domain is not only governed by $\theta_l$ but also $\theta_u$ by using $p(z^u_w | \theta_l, \theta_u)$. The transfer learning is performed by incorporating the topic mixtures of source and target domains in generation of $z^u_w$.

The data generation is described as follows:

For each source document $d_l$:
1. Choose a category by $c_l \sim p(c_l | \rho)$ for each document. Here, $\rho$ denotes a $C$-dimensional multinomial parameter.
2. Draw a topic mixture by a prior density $\theta_l \sim \text{Dir}(\theta_l | c_l, \alpha_l)$ where $\theta_l$ is the parameter of a multinomial distribution for choosing topic $z$ and $\alpha_l$ is the Dirichlet parameter conditioned on category $c$ with $K$ topics.
3. For each of $N$ words $w^l_u$ in the document
   (a) Choose a topic by $z^l_w \sim \text{Mult}(\theta_l, \theta_u)$.
   (b) Choose $w^l_u$ by a multinomial distribution $p(w^l_u | z^l_w, \beta)$ conditioned on $z^l_w$ and a $K \times V$ matrix $\beta = \{\beta_{vu} = p(w^l_u | z^l_w = 1)\}$.

For each target document $d_u$:
1. Draw a topic mixture by the prior density $\theta_u \sim \text{Dir}(\theta_u | \alpha_u)$.
2. For each word $w^u_u$ in the document
   (a) Choose a topic $z^u_w \sim \text{Mult}(\theta_u)$ using the parameter $\theta_u$ of target domain.
   (b) Choose a word $w^u_u$ from $p(w^u_u | z^u_w, \beta)$ conditioned on $z^u_w$ and using the shared parameter $\beta$.

Unlike the TPLSA [29], BTM aims to construct a Bayesian topic model by compensating the uncertainty due to the latent topic $z$ or topic mixtures $\theta$. The topic information in
BTM is drawn from a Dirichlet prior density $\theta_u \sim \text{Dir}(\theta | \alpha)$, is expressed by

$$p(\theta | \alpha) = \frac{\Gamma \left( \sum_{k=1}^{K} \alpha_k \right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \theta_u^{\alpha_u-1} \cdots \theta_K^{\alpha_K-1}$$

(15)

where $\Gamma(\cdot)$ is the gamma function and the parameter $\alpha$ is a $k$-vector with components $\alpha_i > 0$.

Given the parameters $\alpha$ and $\beta$, the joint distribution of observations $\{D_s, D_t\}$ and latent variables $\{z, \theta_l, \theta_u\}$ is yielded as

$$p(w, c, z, \theta | \alpha, \beta, \rho) = \prod_{j=1}^{M} \left[ p(c_j | \rho_j) p(\theta_l | \alpha^l, c_j) \prod_{u=1}^{N^l} p(z^l_u | \theta_l, \theta_u) p(w^l_u | z^l_u, \beta) \right]$$

$$\times \prod_{u=1}^{N^u} \left[ p(\theta_u | \alpha^u) \prod_{v=1}^{N^u} p(z^u_v | \theta_u) p(w^u_v | z^u_v, \beta) \right].$$

(16)

The marginal likelihood is then calculated by

$$p(w, c | \alpha, \beta, \rho) = \int \int \prod_{j=1}^{M} \left[ p(c_j | \rho_j) p(\theta_l | \alpha^l, c_j) \prod_{u=1}^{N^l} p(z^l_u | \theta_l, \theta_u) p(w^l_u | z^l_u, \beta) \right]$$

$$\times \prod_{u=1}^{N^u} \left[ p(\theta_u | \alpha^u) \prod_{v=1}^{N^u} p(z^u_v | \theta_u) p(w^u_v | z^u_v, \beta) \right] d\theta_u d\theta_l.$$

(17)

Due to the marginalization over the priors of source and target topic mixtures $\{\theta_l, \theta_u\}$, we establish the BTM as a regularized model [30] for text classification.

### 3.2 Model Inference and Learning

The exact inference of BTM model parameters does not exist since the marginal likelihood in Eq. (17) is intractable. Similar to model inference using LDA, we employ the VB-EM algorithm and maximize the lower bound of log marginal likelihood

$$\log p(w, c | \alpha, \beta, \rho) \geq E_q[\log p(c, z, \theta, w | \alpha, \beta, \rho)]$$

$$- E_q[\log q(z, \theta)].$$

(18)

A variational distribution $q(z, z, \theta, \theta | \phi, \gamma, \eta)$ over latent variables $\{z, z, \theta, \theta\}$ is used to approximate the posterior distribution $p(z, \theta | w, \alpha, \beta)$, where $\{\phi, \gamma, \eta\}$ are the corresponding variational parameters. Considering the factorized variational inference, the lower bound in Eq. (18) is decomposed as
Maximizing the lower bound is equivalent to minimizing the Kullback-Leibler (KL) distance \[15\] between the variational posterior probability and the true posterior probability \[3\]. Therefore, we obtain the updating equations for variational parameters \{\phi, \varphi, \gamma, \eta\} and BTM parameters \{\alpha, \beta\} by applying VB-EM algorithm. We estimate the optimal variational BTM parameters \{\phi, \varphi, \gamma, \eta\} and BTM parameters \{\alpha, \beta\} to approximate the expectation function of (19) in VB-E step. In the VB-M step, we maximize the expectation function \[L(\gamma; \phi, \varphi, \gamma, \eta)\] with respect to \(\beta \in \{\beta_{v}k\}\) and obtain the optimal parameters

\[
\beta_{v}k \propto \sum_{j=1}^{N_{v}} \sum_{c=1}^{C} \phi_{d_{v}c} \omega_{v}^{c} \sum_{k=1}^{K} \phi_{d_{v}c} \omega_{v}^{c} .
\] (24)

Analogous to \[3\], Dirichlet parameter \(\alpha\) is estimated by the Newton-Raphson algorithm and obtained

\[
\alpha_{t+1} = \alpha_{t} - H(\alpha_{t})^{-1}g(\alpha_{t})
\] (25)

where \(H(\alpha)\) and \(g(\alpha)\) denotes the Hessian matrix and gradient vector of lower bound with respect to \(\alpha = \{\alpha_{v}^{c}, \alpha_{v}^{d}\}\) at iteration \(t\), respectively. \(\Psi(\cdot)\) denotes the digamma function, i.e. the first derivative of log gamma function. The updating of these parameters is performed iteratively until convergence. The trained model parameters are used for text classification. From Eqs. (20)-(24), we can see that the variational parameters \{\phi, \varphi, \gamma, \eta\} and BTM parameters \{\alpha, \beta\} contain the mixed information from source documents \(D_{s}\) as well as target documents \(D_{t}\). Because \(\mathbf{c}\) is observed and \(p(\mathbf{c} | \mathbf{p})\) is independent of the other parameters, the parameters \{\rho_{c}\} can be calculated from \(\mathbf{c}\) using maximum likelihood estimation. Interestingly, the parameters \(\gamma_{v}k\) and \(\eta_{v}k\) are interpreted as the probabilities of using topic \(k\) to generate the source document \(d_{v}\) at category \(c\) and the target document \(d_{v}\), respectively. The topic-dependent word probabilities \(\beta = \{\beta_{v}k\}\) are calculated from the word co-occurrences in source documents and target documents. After all parameters
converged, the parameters \( p(\theta | z) \) are used to assign each target document to a category.

### 3.3 Discriminative Learning for BTM

PLSA and LDA are generative models, which are estimated using the ML method. From the viewpoint of pattern recognition, the discriminative model is generally better than the generative model because classification performance is measured by classification accuracy rather than the likelihood function, which matches observation data with the target model parameters. Discriminative training has been successfully applied in various tasks for machine learning and pattern recognition. To fulfill the Bayes decision rule, the Bayes risk, which is based on the classification error loss function, was minimized to estimate the model [13]. Model discrimination between the target and its competing classes was improved. BTM is estimated by maximizing the likelihood function of source and target documents \( D \) given the document words \( w \) as follows

\[
p(D) = \int p(\theta | \alpha) \prod_{n=1}^{N} \sum_{z_n} p(w_n | \theta) p(z_n | \theta) d\theta \propto \left( \sum_{k=1}^{K} \theta_k \beta_{nk} \right)^{g_{n-d}}
\]

where \( g_{n,d} \) is the number of occurrences of a word \( w_n \) in word set \( w \). In the test session, a test document is classified according to the maximum a posteriori decision rule \( c^* = \arg \max p(c | w_{du}) \). A desirable classification was built with low classification errors for source and target documents. In this study, we attempted to adjust the BTM or re-estimate the model parameters according to the MCE criterion, which is consistent with the metrics of pattern classification in the test session. The DBTM was accordingly, and is shown in Fig. 2. This procedure starts from an initial document model trained by the BTM. In the re-estimation stage, we adjust the BTM parameters, the source documents with the class labels, and the target document with the labels obtained, by the current classifier. The objective function in MCE training is constructed by firstly calculating a misclassification measure expressed by

\[
d_{MCE}(D; \theta, \beta) = -\log p(D | c_i; \theta, \beta) + \log \left( \frac{1}{C-1} \sum_{j \neq i} \exp[\log p(D | c_j; \theta, \beta)] \right)^{1/q}
\]

where the likelihood functions of training documents given the correct/target classes \( c_i \)
and the incorrect/competing classes $c_j$ are involved in Eq. (27). Given the misclassification measure, a general form of loss function is defined as

$$l(d_{MCE}(D; \theta, \beta)) = \frac{1}{1 + \exp(-a d_{MCE}(D; \theta, \beta) + b)}.$$  (28)

The objective function is obtained by the expectation of loss function

$$L_C(\theta, \beta) = E_d[l_{MCE}(D; \theta, \beta)].$$  (29)

Finally, the model parameters are computed by minimizing the expected loss through the descent algorithm

$$(\theta, \beta)_{MCE}^{\text{opt}} = (\theta, \beta)_{MCE}^{\text{opt}} - \varepsilon \nabla l_{MCE}(D; \theta, \beta)_{MCE}.$$  (30)

which is an iterative procedure with learning rate $\varepsilon$ and iteration index $t$. The discriminative model using DBTM is established when the convergence condition is met.

4. EXPERIMENTAL RESULTS

4.1 Data Sets

We evaluated the proposed methods for cross-domain text classification and compared it with the related methods. We conducted the experiments on using 20 News-groups collection. This data set was a public-domain collection, which contained approximately 20,000 newsgroup documents and was partitioned nearly evenly across 20 different newsgroups (about 1,000 documents per newsgroup) [16]. We generated ten different data sets from four top-categories (comp, sci, rec and talk) for evaluating classification, which contained 16,015 newsgroup documents and 61,188 words. The term frequency was used as the features. As referred to [29], the dataset was arranged in a hierarchical structure. The sub-categories of a top-category were split into two sets; one was for training and the other was for testing. The training and test data were different but in the related domains. For example, the top-category of “science” in source documents covered the topics of “crypt” and “electronics” while the target documents of the same class also covered the topics of “med” and “space”. Such a split is similar to what used in [25, 29].

4.2 Performance on Different Datasets

The AUC [17] and the $F$-measure [22] were reported to evaluate different methods. We compare the effectiveness of NBC, support vector machine (SVM), latent Dirichlet allocation (LDA), topic-bridge PLSA (TPLSA) and the proposed BTM. The term frequency was used as the features. Using TPLSA and BTM methods, we adopted 50 topics and run 100 iterations in VB-EM procedure. In TPLSA, we set parameters $\lambda = 0.5$, $\nu_1 = 50$, and $\nu_2 = 15$. TPLSA and BTM conducted transfer learning using source and target
Table 1. Comparison of AUC/F-measures for different methods.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>NBC</th>
<th>SVM</th>
<th>LDA</th>
<th>TPLSA</th>
<th>BTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>rec vs talk</td>
<td>0.791/0.760</td>
<td>0.770/0.715</td>
<td>0.805/0.823</td>
<td>0.909/0.901</td>
<td><strong>0.942/0.925</strong></td>
</tr>
<tr>
<td>rec vs sci</td>
<td>0.783/0.794</td>
<td>0.737/0.726</td>
<td>0.820/0.869</td>
<td>0.920/0.881</td>
<td><strong>0.914/0.912</strong></td>
</tr>
<tr>
<td>comp vs talk</td>
<td>0.929/0.900</td>
<td>0.876/0.832</td>
<td>0.933/0.912</td>
<td>0.945/0.920</td>
<td><strong>0.966/0.948</strong></td>
</tr>
<tr>
<td>comp vs sci</td>
<td>0.776/0.775</td>
<td>0.699/0.694</td>
<td>0.825/0.881</td>
<td>0.903/0.917</td>
<td><strong>0.913/0.939</strong></td>
</tr>
<tr>
<td>comp vs rec</td>
<td>0.892/0.841</td>
<td>0.825/0.745</td>
<td>0.887/0.867</td>
<td>0.910/0.870</td>
<td><strong>0.926/0.882</strong></td>
</tr>
<tr>
<td>sci vs talk</td>
<td>0.770/0.755</td>
<td>0.744/0.710</td>
<td>0.806/0.846</td>
<td>0.915/0.892</td>
<td><strong>0.902/0.907</strong></td>
</tr>
</tbody>
</table>

Average 0.823/0.804 0.775/0.737 0.846/0.866 0.917/0.897 **0.927/0.919**

For three categories classification, we tested and compared our model BTM with NBC and TPLSA. Fig. 3 shows the results. Again, these results show that BTM outperforms NBC and TPLSA. The experimental results shown in the preceding Table 1 and Fig. 3 are based on the two or three categories. Next, the models are extended to four categories (sci vs rec vs talk vs comp). In the set of experiments, BTM achieves the F-measure of 0.584, offering a 3.1% improvement with respect to TPLSA (0.553), and an 11.6% improvement with respect to NBC (0.468).

Next, we examine the effect of discriminative training in cross-domain document classification. In the implementation, DBTM adopted the parameters $\eta = 1$, $a = 1$, $b = 0$, and $\epsilon = 0.0001$. Table 2 lists the experimental results of BTM and DBTM in cases of datasets “rec vs talk” and “comp vs sci”. The discriminative training did further improve the classification performance of test data. The AUC of DBTM using “rec vs talk” and “comp vs sci” was 0.947 and 0.917, respectively, which is higher than that of BTM.
Table 2. Comparison of AUC/F-measures of BTM and DBTM.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>BTM</th>
<th>DBTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>rec vs talk</td>
<td>0.942/0.925</td>
<td>0.947/0.932</td>
</tr>
<tr>
<td>comp vs sci</td>
<td>0.913/0.939</td>
<td>0.917/0.943</td>
</tr>
</tbody>
</table>

Table 3. Top: The most commonly occurring words in some of the topics inferred from the 20 Newsgroups training data by TPLSA. Bottom: Some of the topics inferred by the BTM model. Each column represents a single topic, and words appear in order of frequency of occurrence. Content words are in bold.

<table>
<thead>
<tr>
<th>Class</th>
<th>Comp</th>
<th>Sci</th>
</tr>
</thead>
<tbody>
<tr>
<td>TPLSA</td>
<td>data</td>
<td>key</td>
</tr>
<tr>
<td></td>
<td>sound</td>
<td>to</td>
</tr>
<tr>
<td>for</td>
<td>that</td>
<td>the</td>
</tr>
<tr>
<td>modem</td>
<td>microsoft</td>
<td>encryption</td>
</tr>
<tr>
<td>times</td>
<td>ms</td>
<td>and</td>
</tr>
<tr>
<td>cd</td>
<td>of windows</td>
<td>will</td>
</tr>
<tr>
<td>that</td>
<td>product</td>
<td>escrow</td>
</tr>
<tr>
<td>hardware</td>
<td>rate</td>
<td>that</td>
</tr>
<tr>
<td>port</td>
<td>dos</td>
<td>microsoft</td>
</tr>
<tr>
<td></td>
<td>for</td>
<td>ms</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dos</td>
</tr>
<tr>
<td>BTM</td>
<td>disk</td>
<td>the</td>
</tr>
<tr>
<td></td>
<td>copy</td>
<td>to</td>
</tr>
<tr>
<td></td>
<td>hard</td>
<td>law</td>
</tr>
<tr>
<td></td>
<td>user</td>
<td>of</td>
</tr>
<tr>
<td></td>
<td>the</td>
<td>government</td>
</tr>
<tr>
<td></td>
<td>program</td>
<td>law</td>
</tr>
<tr>
<td></td>
<td>software</td>
<td>to</td>
</tr>
<tr>
<td></td>
<td>disks</td>
<td>will</td>
</tr>
<tr>
<td></td>
<td>install</td>
<td>will</td>
</tr>
<tr>
<td>os</td>
<td>have</td>
<td>com</td>
</tr>
</tbody>
</table>

(0.942 and 0.913). The improvements of F-measure also obtained.

To investigate the effect of different sizes of training data, we compare the F-measures of TPLSA and BTM by changing the percentage of training data size. The dataset “rec vs talk” was adopted. As shown in Fig. 4, the F-measure curve of BTM is higher than that of TPLSA. These two topic models were not sensitive to the change of data size.

4.3 Analysis of Text Modeling

In addition to comparing AUC or F-measure, it is instructive to look at the inferred topics. We also interest to examine the discovered topics and their association with class labels. Table 3 shows the most commonly occurring words assigned to a selection of domain-specific topics extracted from the Comp vs. Sci dataset. Each topic is illustrated with the top 10 words most likely to be generated conditioned on the topic. The topic inferred using TPLSA contain more function words than BTM model, such as “of”, “to” and “the”. Besides, BTM model selects more terms related to “science” and “computer” than TPLSA. For example, in “computer”, some content words (such as “speed”, “driver”, “memory”) are actually the components of the popular words in this area as shown in the BTM model. Some extremely related words (such as “disk”), ranked very high in
Table 4. Data arrangements for multiple source tasks.

<table>
<thead>
<tr>
<th>Training Data $D_s$</th>
<th>$D^1_s$</th>
<th>$D^2_s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>rec vs talk</td>
<td>rec.autos</td>
<td>rec.motorcycles</td>
</tr>
<tr>
<td></td>
<td>talk.politics.guns</td>
<td>talk.politics.misc</td>
</tr>
<tr>
<td>comp vs sci</td>
<td>comp.graphics</td>
<td>comp.os.ms-window-misc</td>
</tr>
<tr>
<td></td>
<td>sci.crypt</td>
<td>sci.electronics</td>
</tr>
</tbody>
</table>

Table 5. Comparison among LDA, TPLSA and BTM on a multiple source learning task.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>LDA</th>
<th>TPLSA</th>
<th>BTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>rec vs talk</td>
<td>0.853</td>
<td>0.914</td>
<td>0.928</td>
</tr>
<tr>
<td>comp vs sci</td>
<td>0.896</td>
<td>0.931</td>
<td>0.954</td>
</tr>
</tbody>
</table>

BTM, are absent in TPLSA’s top word list. In the “science”, we can find similar phenomena as well. The result shows that our method can effectively identify the correlations between domain-specific features from different domains.

4.4 Performance on Multiple Source Domains

Here, we conduct experiments to show that the BTM can also work on multiple source domains. We evaluate TPLSA and BTM methods on the problem with 2 sources and 1 target. For comparison, LDA was carried out as a baseline system, which did not involve transfer learning. The single source domain $D_s$ in datasets “rec vs talk” and “comp vs sci” was further divided into two source domains $\{D^1_s, D^2_s\}$, which are described in Table 4 and the target domain $D_t$ was not changed as arranged in [18]. Table 5 shows the results in terms of $F$-measure. These results show that BTM outperforms the baseline method (LDA) and TPLSA on the tasks with multiple source domains.

5. CONCLUSION

This paper presented the cross-domain discriminative training approach to estimate the Bayesian topic document model over different domains. We developed a statistical latent topic cross-domain model, which was jointly estimated from labeled and unlabeled documents. A bridged topic model was established for cross-domain text classification. A variational inference algorithm was formulated to find BTM parameters via a VB-EM estimation procedure. Moreover, the minimum classification error criterion was integrated to perform discriminative training by combing the likelihood functions of target models and competing models. The experiments on document classification using 20 Newsgroups dataset validated the proposed method in terms of $F$-measure over different training data sizes and different methods.

ACKNOWLEDGMENT

The authors would like to thank Prof. Zhou and three anonymous reviewers for their valuable comments, which improved the presentation of this paper.
REFERENCES


Meng-Sung Wu (吳孟淞) received his Ph.D. degree in Department of Computer Science and Information Engineering from National Cheng Kung University in 2009. In 2009, he joined the Machine Learning Lab at NCKU as a Postdoctoral Researcher. From 2010 to 2011, he served as Postdoctoral Fellow at the Insti-
MENG-SUNG WU

Institute of Information Science, Academia Sinica, Taipei, Taiwan. Now he is working in Computational Intelligence Technology Center, Industrial Technology Research Institute. His current research interests include text and language processing, information retrieval and classification, big data analysis, knowledge discovery from text. He is a member of IEEE and ACM.