Supervised labeled latent Dirichlet allocation for document categorization

Ximing Li · Jihong Ouyang · Xiaotang Zhou · You Lu · Yanhui Liu

Published online: 25 November 2014
© Springer Science+Business Media New York 2014

Abstract Recently, supervised topic modeling approaches have received considerable attention. However, the representative labeled latent Dirichlet allocation (L-LDA) method has a tendency to over-focus on the pre-assigned labels, and does not give potentially lost labels and common semantics sufficient consideration. To overcome these problems, we propose an extension of L-LDA, namely supervised labeled latent Dirichlet allocation (SL-LDA), for document categorization. Our model makes two fundamental assumptions, i.e., *Prior* 1 and *Prior* 2, that relax the restriction of label sampling and extend the concept of topics. In this paper, we develop a Gibbs expectation-maximization algorithm to learn the SL-LDA model. Quantitative experimental results demonstrate that SL-LDA is competitive with state-of-the-art approaches on both single-label and multi-label corpora.

Keywords Supervised · Topic modeling · Latent Dirichlet allocation · Multi-label classification

X. Li · J. Ouyang (✉) · X. Zhou · Y. Lu · Y. Liu
College of Computer Science and Technology,
Jilin University, Changchun, China
e-mail: ouyj@jlu.edu.cn

X. Li
e-mail: liximing86@gmail.com

X. Li · J. Ouyang · X. Zhou · Y. Lu · Y. Liu
Key Laboratory of Symbolic Computation and
Knowledge Engineering of Ministry of Education,
Jilin University, Changchun, China

1 Introduction

Recently, considerable attention has been focused on topic modeling approaches [8]. The original goals of such methods were (1) to obtain a brief description of document collection for basic tasks [1] such as classification, clustering, and dimension reduction, and (2) to use the concept of latent topics to capture the semantics behind documents. Latent Dirichlet allocation (LDA) [5] is acknowledged as the most successful topic model. It simulates the generative process of a corpus, where each document is composed of latent topics, and each topic is described by a multinomial distribution over words. To control the capacity of the model parameters and avoid the over-fitting problem, a Dirichlet prior is given over all topics beyond the corpus.

Most approaches to topic modeling are unsupervised, and LDA is no exception. Unsupervised LDA neglects supervised information, resulting in some significant messages being wasted. For document classification tasks [5, 6, 14], LDA commonly acts as an upstream method for dimension reduction before various classifiers are executed, e.g., LDA with support vector machines (SVMs). Intuitively, some discriminative features must be lost when LDA transforms the original word distribution into the latent topic distribution.

The investigation of supervised LDA models faces the challenge of incorporating supervision into the learning procedure. Recently, some modifications have been developed (e.g., supervised LDA (sLDA) [4], discriminative LDA (DiscLDA) [12] and maximum entropy discrimination LDA (MedLDA) [23, 24] for single-label classification; labeled LDA (L-LDA) [15], Dirichlet process with mixed random measures (DP-MRM) [11], Prior-LDA (Pr-LDA), and Dependency-LDA (Dep-LDA) [17] for multi-label classification).

To the best of our knowledge, L-LDA was the first supervised LDA model that could be applied to multi-label corpora. L-LDA defines a 1-1 correspondence between labels and topics, and constrains each document to its pre-assigned label set. The empirical results reported in [15] show that, under certain conditions, L-LDA can achieve performance that is competitive with state-of-the-art discriminative methods, e.g., SVMs [13]. However, it suffers from two problems: (1) L-LDA is over-focused on pre-assigned labels for each document, resulting in worse performance for corpora with small *cardinality* values (i.e., the average number of labels per document), and even degenerating to the mixture of unigrams model for single-label corpora; (2) from a generative perspective, L-LDA lacks a mechanism to capture potentially lost labels and common semantics.

In this paper, we consider these two problems, and then propose a Supervised L-LDA (SL-LDA) model for both single- and multi-label document categorization. Our model makes two fundamental assumptions beyond those used in L-LDA: *Prior* 1 provides a label threshold to highlight the pre-assigned labels, instead of restricting them; *Prior* 2 extends the topics concept to cover the semantics of lost labels and common topics. We develop a Gibbs Expectation-Maximization (Gibbs-EM) algorithm to infer the SL-LDA model. Naturally, as a supervised model, SL-LDA can directly predict the response labels for the test documents.

Extensive experiments are conducted to evaluate the proposed SL-LDA model. Several traditional approaches are chosen as performance baselines, including: 1) for the single-label case, four LDA-based approaches, i.e., LDA-SVM, sLDA, DiscLDA, and MedLDA; 2) for the multi-label case, three supervised LDA models, i.e., L-LDA, Pr-LDA, and Dep-LDA, and the state-of-the-art discriminative algorithm SVMs [13, 21]. The empirical results show that SL-LDA achieves a performance level that is competitive with the state-of-the-art methods. Some important notation used in this paper is summarized in Table 1.

The remainder of this paper is organized as follows. Section 2 describes the proposed SL-LDA model in detail. Section 3 presents the parameter estimation and inference methods for SL-LDA. In Section 4, we introduce some related studies on supervised topic models. Sections 5 and 6 present and discuss the empirical results for single- and multi-label corpora, respectively. Finally, our conclusions and suggestions for future work are given in Section 7.

2 Proposed model

This section introduces the novel SL-LDA model. We begin by reviewing the L-LDA model.

Table 1 Notation descriptions

Notation	Description
D	number of documents
K	total number of topics
K_t	number of labels
K_h	number of hidden topics in SL-LDA
V	number of words
$\overrightarrow{\Lambda_d}$	the topic presence/absence indicator w.r.t document d
θ^*	the label threshold in SL-LDA
$\overrightarrow{\theta_d}$	the multinomial distribution of topics w.r.t document d
Θ	the multinomial distribution of topics w.r.t all documents
$\overrightarrow{\alpha}$	the Dirichlet prior for each $\overrightarrow{\theta_d}$
$\overrightarrow{\phi_k}$	the multinomial distribution of words w.r.t topic k
Φ	the multinomial distribution of words w.r.t all topics
$\overrightarrow{\beta}$	the Dirichlet prior for each $\overrightarrow{\phi_k}$
y_d	the pre-assigned label set for document d
K_d	the value of $\|y_d\|$

2.1 L-LDA

L-LDA [15] is a supervised topic model for describing labeled corpora. Similar to LDA, L-LDA represents documents as multinomial distributions over topics, where each topic is represented by a multinomial distribution over words. Additionally, L-LDA defines a 1-1 correspondence between the labels (tagged by human-being) and topics. Each document d is constrained to be described by its pre-assigned label set $y_d \subseteq \{1, 2, \cdots, K_t\}$.

L-LDA is formalized as follows: For each topic k, generate the topic-word distribution $\overrightarrow{\phi_k}$, drawn from Dirichlet prior $\overrightarrow{\beta}$. However, for each document d, L-LDA restricts the multinomial distribution $\overrightarrow{\theta_d}$ to topics within y_d. Towards this requirement, the topic presence/absence indicator $\overrightarrow{\Lambda_d} = \left(l_1, l_2, \cdots, l_{K_t}\right)$ is generated via a Bernoulli distribution η_k, where $l_k \in \{0, 1\}$. Define $y_d = \left\{k|\Lambda_{d,k} = 1\right\}$. The document-label projection matrix L_d [15] is used to project the topic Dirichlet prior $\overrightarrow{\alpha}$ to a presence topic prior $\overrightarrow{\alpha_d} = (\alpha_{d,l_1}, \alpha_{d,l_2}, \cdots, \alpha_{d,l_{K_d}})^T$. Then, the topics and words can be sampled as in the traditional LDA model.

In summary, L-LDA assumes the following generative process for a labeled corpus W:

1. For each topic k
   (a) Choose $\overrightarrow{\phi_k} = \left(\phi_{k,1}, \phi_{k,2}, \cdots, \phi_{k,V}\right)^T \sim Dirichlet\left(\overrightarrow{\beta}\right)$
2. For each document d in the corpus W
   (a) For each topic k
      (i) Choose $\Lambda_{d,k} \in \{0, 1\} \sim Bernoulli\,(\eta_k)$

(b) Choose $\overrightarrow{\alpha_d} = L_d \times \overrightarrow{\alpha}$

(c) Choose $\overrightarrow{\theta_d} = (\theta_{d,l_1}, \theta_{d,l_2}, \cdots, \theta_{d,l_{K_d}})^T \sim Dirichlet\left(\overrightarrow{\alpha_d}\right)$

(d) For each of the N_d words $w_{d,n}$

   (i) Choose a topic $z_{d,n} \in \left\{y_{d,1}, y_{d,2}, \cdots, y_{d,K_d}\right\} \sim Multinomial\left(\overrightarrow{\theta_d}\right)$

   (ii) Choose a word $w_{d,n}$ from $p\left(w_{d,n}| z_{d,n}, \Phi\right)$, a multinomial probability conditioned on the topic $z_{d,n}$

2.2 SL-LDA

In reviewing the L-LDA model, we found that it suffers from two problems in terms of supervised tasks. First, the assumption that constrains each document to be sampled from its own label set y_d is too strong for corpora with small *cardinality*. For single-label corpora, in particular, L-LDA will degenerate to the simple mixture of unigrams model. This undermines the advantage of capturing the latent semantics of topic modeling, and leads to worse classification performance. Second, because L-LDA defines a 1-1 correspondence between labels and topics, its number of topics is equivalent to the true number of labels in the corpus. From a generative perspective, L-LDA ignores the hidden labels (i.e., lost labels), which might be neglected during manual processing. More importantly, this framework of L-LDA lacks a mechanism to cover the common semantics. For example, the words "news" and "report", which express common semantics in newsgroup collections, might occur in most documents. In the context of L-LDA, such common words are forcibly assigned a pre-defined label. Unfortunately, these words are less discriminative, so this is harmful to the classification.

To overcome these problems, we propose a novel supervised model. SL-LDA makes two fundamental assumptions beyond those of L-LDA, namely *Prior* 1 and *Prior* 2. The aim of *Prior* 1 is to relax the restriction of label sampling. Following the intuition that a document involves a wide range of labels, but that only the main labels will be tagged (e.g., given a document d with label proportion [0.1, 0.2, 0.2, 0.5], its label set might be manually given as $y_d = \{4\}$), *Prior* 1 supposes that each document d samples from all labels and gives the dominant weights, i.e., the label threshold θ^*, for labels in y_d. Under this assumption, SL-LDA will never degenerate to a simpler model. *Prior* 2 extends the concept of topics to handle the problem of semantic loss. It assumes that each document d is represented by K topics, consisting of K_t observed labels and K_h hidden topics. The hidden topics are used to cover the potentially lost labels and common semantics, which are represented by the common words. Consequently, less discriminative common words, such as "news" and "report" in newsgroup collections, contribute to the hidden topics rather than the pre-defined labels. Although *Prior* 2 seems straightforward, it provides significant benefits to classification tasks.

Note that in the structure of SL-LDA the observed labels and hidden topics are on the same level. Thus, in this paper, we do not differentiate between the concepts of "label" and "topic". *Prior* 1 and *Prior* 2 are summarized as follows:

– *Prior* 1: each document d samples from all topics; the labels in y_d are given dominative proportions, defined by θ^*.
– *Prior* 2: there are a total of K topics, consisting of K_t observed labels and K_h hidden topics.

Formally, SL-LDA generates the topic-word distribution $\overrightarrow{\phi_k}$ as in L-LDA. For each document d, it first samples the topic presence/absence indicator $\overrightarrow{\Lambda_d}$ under the observed labels, and then generates the document-topic distribution $\overrightarrow{\theta_d}$ via the following conditional function:

$$fun(\overrightarrow{\alpha}, y_d, \theta^*, K_d) = \begin{cases} \theta_{d,i} = \frac{\theta^*}{K_d} & i \in y_d \\ \\ \sim Dir'(\overrightarrow{\alpha}, y_d) & otherwise \end{cases} \tag{1}$$

where $\theta^* \leq 1$. For each document d, $\sum_{i \in y_d} \theta_{d,i} = \theta^*$ and $\theta_{d,i} = \theta_{d,j}, \ \forall i, j \ \in y_d$.; $Dir'(\overrightarrow{\alpha}, y_d)$ is a pseudo-Dirichlet distribution with the following Probability Density Function (PDF)[1]:

$$Dir'(\overrightarrow{\alpha}, y_d) = \frac{\Gamma\left(\sum_{k \notin y_d} \alpha_k\right)}{(1-\theta^*)^{\sum_{k \in y_d} \alpha_k} \times \prod_{k \notin y_d} \Gamma(\alpha_k)} \prod_{k \notin y_d} \theta_k^{\alpha_k - 1}.$$

$$\sum_{k \notin y_d} \theta_k = 1 - \theta^*.$$

where $\Gamma(x)$ is the Gamma function.

The generative process of SL-LDA can be summarized as follows:

1. For each topic k

   (a) Choose $\overrightarrow{\phi_k} = \left(\phi_{k,1}, \phi_{k,2}, \cdots, \phi_{k,V}\right)^T \sim Dirichlet\left(\overrightarrow{\beta}\right)$

2. For each document d in the corpus W

   (a) For each existing topic k

      (i) Choose $\Lambda_{d,k} \in \{0, 1\} \sim Bernoulli\,(\eta_k)$

   (b) Choose $\overrightarrow{\theta_d} = (\theta_{d,1}, \theta_{d,2}, \cdots, \theta_{d,K})^T \sim f(\overrightarrow{\alpha}, y_d, \theta^*, K_d)$

[1] In fact, $Dir'(\overrightarrow{\alpha}, y_d)$ is a pseudo-PDF, because $\int_{-\infty}^{\infty} Dir'(\overrightarrow{\alpha}, y_d) = 1 - \theta^*$. Here, it is used to sample the absence topic components in $\overrightarrow{\Lambda_d}$.

(c) For each of the N_d words $w_{d,n}$

  (i) Choose a topic $z_{d,n} \in \{k_1, k_2, \cdots, k_K\} \sim Multinomial\left(\overrightarrow{\theta_d}\right)$

  (ii) Choose a word $w_{d,n}$ from $p\left(w_{d,n} | z_{d,n}, \Phi\right)$, a multinomial probability conditioned on the topic $z_{d,n}$

Discussion of parameter θ^* The parameter θ^* is a crucial threshold function that determines the degree of attention of the marked labels. We suggest the tuned interval $\theta^* \in [0.5, 0.8]$ following the intuition that: (1) a document is assigned a label (labels), provided that at least 50 % is focused on this label (labels); (2) a document is over-focused on the marked label (labels) if θ^* approaches 1.

In Sections 5.2.3 and 6.4, we study θ^* experimentally. The results show that the optimum performance is achieved when θ^* is within the suggested interval.

2.3 Evidence for SL-LDA

Given the parameters Φ and Θ , the joint probability of the labeled corpus W and a set of topic assignments $\overrightarrow{z}$ is:

$$P\left(W, \overrightarrow{z} | \Theta, \Phi\right) = \prod_{d=1}^{D} \prod_{k=1}^{K} \prod_{v=1}^{V} \theta_{d,k}^{N_{d,k}} \phi_{k,v}^{N_{k,v}}$$

$$\left(\theta_{d,k} = \frac{\theta^*}{K_d}, \quad if\ k \in y_d\right) \tag{2}$$

where $N_{d,k}$ is the number of times that the topic k occurs in document d; and $N_{k,v}$ is the number of times that the word v has been assigned to topic k.

SL-LDA places a Dirichlet prior over Φ:

$$P\left(\Phi | \overrightarrow{\beta}\right) = \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma\left(\beta_v\right)} \prod_{v=1}^{V} \phi_v^{\beta_v - 1} \tag{3}$$

and a pseudo-Dirichlet prior over Θ:

$$P\left(\Theta | \overrightarrow{\alpha}\right) = \prod_{d=1}^{D} \frac{\Gamma\left(\sum_{k \notin y_d} \alpha_k\right)}{(1-\theta^*)^{\sum_{k \in y_d} \alpha_k} \times \prod_{k \notin y_d} \Gamma\left(\alpha_k\right)} \prod_{k \notin y_d} \theta_k^{\alpha_k - 1} \tag{4}$$

Given the hyper-parameters $\overrightarrow{\alpha}$ and $\overrightarrow{\beta}$, we obtain the evidence for a labeled corpus W by combining (2), (3) and (4):

$$P\left(W | \Theta, \Phi\right) = \sum_{z} \left(C^* \prod_{d=1}^{D} \frac{\Gamma\left(\sum_{k \notin y_d} \alpha_k\right)}{\prod_{k \notin y_d} \Gamma\left(\alpha_k\right)} \frac{\prod_{k \notin y_d} \Gamma\left(\alpha_k + N_{d,k}\right)}{\Gamma\left(\sum_{k \notin y_d} \left(\alpha_k + N_{d,k}\right)\right)} \right.$$
$$\left. \prod_{k=1}^{K} \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma\left(\beta_v\right)} \frac{\prod_{v=1}^{V} \Gamma\left(\beta_v + N_{k,v}\right)}{\Gamma\left(\sum_{v=1}^{V} \left(\beta_v + N_{k,v}\right)\right)} \right) \tag{5}$$

where variable C^* is defined, for simplification, as follows:

$$C^* = \frac{(\theta^*)^{\sum_{d=1}^{D} \sum\limits_{k \in y_d} N_{d,k}}}{\prod_{d=1}^{D} K_d^{\sum_{k \in y_d} N_{d,k}}} \left(1 - \theta^*\right)^{\sum_{d=1}^{D} \sum\limits_{k \notin y_d} N_{d,k}}$$

3 Estimation and inference

In this section, we describe the process of parameter estimation and inference with respect to SL-LDA.

3.1 Estimation for hyper-parameters

We develop a Gibbs-EM algorithm to learn the hyper-parameters $\overrightarrow{\alpha}$ and $\overrightarrow{\beta}$ in SL-LDA. Gibbs sampling [2, 19] is a popular method for approximate learning in high-dimension models. It imitates the high-dimension probability distribution given by the stationary state of Markov chain Monte Carlo (MCMC) chains. We first use Gibbs sampling to approach the expectation of topic assignments $\overrightarrow{z}$: $E\left[P\left(\overrightarrow{z} | W, \overrightarrow{\alpha}, \overrightarrow{\beta}\right)\right]$, and then maximize the likelihood of (5) to estimate the hyper-parameters. Repeating this procedure until convergence, we obtain the optimized values of $\overrightarrow{\alpha}^M$ and $\overrightarrow{\beta}^M$. The Gibbs-EM algorithm is summarized as follows:

1. Initialize topics $\overrightarrow{z}^0$ and hyper-parameters $\overrightarrow{\alpha}^0$ and $\overrightarrow{\beta}^0$
2. For i=1,2,$\cdots$
   (a) E-step: sample for $\overrightarrow{z}^i$ from $P\left(\overrightarrow{z} | W, \overrightarrow{\alpha}, \overrightarrow{\beta}\right)$ by Gibbs sampling.
   (b) M-step: maximize $\log P\left(W, \overrightarrow{z}^i | \overrightarrow{\alpha}, \overrightarrow{\beta}\right)$ w.r.t $\overrightarrow{\alpha}^i$ and $\overrightarrow{\beta}^i$ using the fixed point method

   Until convergence

In the E-step, we sample each topic assignment z_t alternately from the distribution determined by all other topic assignments. Define the symbol $\widetilde{z_t}$ as a set of topic sequences that excludes the current variable z_t. The conditional probability for z_t is then:

$$P\left(z_t = k^* | \widetilde{z_t}, W, \overrightarrow{\alpha}, \overrightarrow{\beta}\right) \propto$$
$$\begin{cases} \dfrac{\theta^*}{K_d} \times \dfrac{N_{\neg k^*, w_t} + \beta_{w_t}}{N_{\neg k^*} + \sum\limits_{v=1}^{V} \beta_v} & if\ k^* \in y_{d_t} \\ (1-\theta^*) \times \dfrac{N_{\neg d_t, k^*} + \alpha_{k^*}}{N_{\neg d_t} + \sum\limits_{k \neq y_{d_t}} \alpha_k} \times \dfrac{N_{\neg k^*, w_t} + \beta_{w_t}}{N_{\neg k^*} + \sum\limits_{v=1}^{V} \beta_v} & else \end{cases} \tag{6}$$

where w_t is the word corresponding to z_t; d_t is the document containing w_t; and variables with the subscript "$\neg$" should have 1 subtracted, e.g., $N_{\neg k^*} = N_{k^*} - 1$.

The M-step uses the fixed point method. Using the samples $\overrightarrow{z}$ obtained in the E-step, we update the hyper-parameter α_k^{i+1} as follows:

$$\alpha_k^{i+1} = \alpha_k^i \times \frac{\sum_{d=1}^{D}\left(\Psi\left(N_{d,k}^i + \alpha_k^i\right) - \Psi\left(\alpha_k^i\right)\right)}{\sum_{d=1}^{D}\left(\Psi\left(\sum_{k\neq y_d}\left(N_{d,y_d}^i + \alpha_k^i\right)\right) - \Psi\left(\sum_{k\neq y_d}\alpha_k^i\right)\right)} \quad \left(N_{d,k}^i = 0,\ if\ k \in y_d\right) \tag{7}$$

and for β_v^{i+1}

$$\beta_v^{i+1} = \beta_v^i \times \frac{\sum_{v=1}^{V}\left(\Psi\left(N_{k,v}^i + \beta_v^i\right) - \Psi\left(\beta_v^i\right)\right)}{\sum_{d=1}^{D}\left(\Psi\left(N_k^i + \sum_{v=1}^{V}\beta_v^i\right) - \Psi\left(\sum_{v=1}^{V}\beta_v^i\right)\right)} \tag{8}$$

where $\Psi(x)$ is the Digamma function, i.e., the logarithmic derivative of the Gamma function.

Consistency over multiple runs In terms of multiple runs, the results of Gibbs sampling might be inconsistent because of the random sampling procedure [9]. This problem persists in our model, but it is almost insignificant for the classification tasks described here. Note that the topics in SL-LDA are composed of observed labels and hidden topics. Given a set of training data, the observed labels are fixed, so they must be consistent. However, because the hidden topics are used to collect common words, they should contribute to the classification regardless of whether they are consistent.

3.2 Inference for unlabeled documents

Let d' be a document from the testing corpus W' and $U = \left\{\overrightarrow{z}, \overrightarrow{w}\right\}$ be a stationary MCMC state for the training corpus W.

We employ Gibbs sampling to infer the unlabeled document d' by estimating the posterior distribution of topic assignments [7]: $P\left(z'|\overrightarrow{w}_{d'}, U, \overrightarrow{\alpha}^M, \overrightarrow{\beta}^M\right)$, where $\overrightarrow{w}_{d'}$ is the vector of d', and $\overrightarrow{\alpha}^M, \overrightarrow{\beta}^M$ are obtained during the parameter estimation procedure. As the test document is unlabeled, the update rule is as follows:

$$P\left(z'_t = k^*|\widetilde{z'}_t, \overrightarrow{w}_{d'}, U, \overrightarrow{\alpha}^M, \overrightarrow{\beta}^M\right) \propto \frac{N_{\neg d',k^*} + \alpha_{k^*}^M}{N_{\neg d'} + \sum_{k=1}^{K}\alpha_k^M} \times \frac{N_{k^*,w'_t} + N'_{\neg k^*,w'_t} + \beta_{w'_t}^M}{N_{k^*} + N'_{\neg k^*} + \sum_{v=1}^{V}\beta_v^M} \tag{9}$$

where w'_t is the *t-th* word in $w_{d'}$; $N'_{k^*,v}$ is the number of times that v has been generated by topic k^*.

Finally, the topic distribution of document d' is estimated as follows:

$$\theta_{d',k} = \frac{N_{d',k} + \alpha_k^M}{\sum_{i=1}^{K}\left(N_{d',i} + \alpha_i^M\right)} \tag{10}$$

where $N_{d',i}^*$ is the number of times that topic i has occurred in document d'.

4 Related work

4.1 Topic model

A number of variants of LDA have been developed for supervised cases. Representative single-label corpora models include sLDA [4], which captures document labels as a classification response, DiscLDA [12], where documents are associated with labels and topic mixtures, and MedLDA [23, 24], which combines maximum margin technology and LDA. In terms of multi-label corpora, L-LDA [15] was the first supervised topic model. The authors of [17] developed the equivalent Flat-LDA, and further extended this model to Pr-LDA and Dep-LDA via observations of label frequency and label dependency, respectively. Other modifications for multi-label corpora include DP-MRM [11] and Partially LDA [16].

The proposed SL-LDA builds upon the L-LDA model by introducing a topic threshold θ^* and extending the concept of topics. In a sense, DiscLDA also extends the topics, as it represents documents by labels and topic mixtures. However, DiscLDA appears to be more complicated than our model. Several models that consider common topics have been explored, e.g., the Cluster-based Topic Model [3] and Multi-Grain Cluster Topic Model [20]. However, as unsupervised models, these cannot be directly applied to classification.

The state-of-the-art Pr-LDA and Dep-LDA models can also be deemed as extensions of L-LDA. Pr-LDA assumes that there is a corpus-wide distribution with respect to the label occurrence frequency in the corpus. A document's topic priors are generated by this frequency distribution, instead of using the same prior. Based on Pr-LDA, Dep-LDA further introduces a topic level beyond the label level to capture label dependency. Although these two models have significantly improved multi-label document classification [17], the two problems of L-LDA still exist. More importantly, these models (particularly Dep-LDA) are complex and contain too many parameters. In terms of their application to different corpora, tuning the parameters might be time-consuming. In contrast, our model involves a simpler construction and fewer parameters, and so is straightforward and can be easily controlled. Another advantage of SL-LDA is that it can be applied to both single- and multi-label corpora. This leads to better scalability in practice.

4.2 Computational complexity

We now discuss the computational complexity of training various supervised topic models. In the original papers, these models were trained using different inference algorithms, and so we provide a descriptive comparison. The traditional LDA model is used as the baseline.

As SL-LDA does not change the basic construction of the traditional LDA model, its complexity will be the same as that of LDA. All single-label models perform extra computations, e.g., calculating the normalization factor in sLDA, learning the transition matrix in DiscLDA, and solving the dual problem in MedLDA. In terms of multi-label models, it is clear that L-LDA is as efficient as LDA, but Pr-LDA and Dep-LDA must compute the upper-topic distribution over the labels. In particular, proper inference using Dep-LDA is time-consuming, because it repeatedly computes the expensive Gamma function. Thus, we argue that the proposed SL-LDA model is more efficient than most of these related supervised topic models. Empirical tests are reported in the following sections.

5 Evaluation on single-label setting

For a single-label corpus ($K_d = 1$), we evaluate both the text modeling and document classification performance of SL-LDA. Experiments are run on the balanced Newsgroups[2] collection, which consists of 19,997 documents in 20 related categories (i.e., $K_t = 20$). By convention, we remove stop words in the standard list[3] and words that occur only once in the corpus.

In the experiments, the parameter θ^* is tuned to a value in the set {0.5, 0.6, 0.7, 0.8}. Other parameters are set as follows: 1) hyper-parameters $\vec{\alpha}$ and $\vec{\beta}$ are initialized to $50/K$ and 0.1, respectively; 2) maximum number of iterations is 50, and the termination precision is 1×10^{-4}; 3) Gibbs sampling uses a burn-in of 500 iterations in the E-step.

5.1 Text modeling

We conducted a simple text modeling experiment to examine the topic structures given by SL-LDA. We fitted the SL-LDA model to the Newsgroups corpus, where $\theta^* = 0.5$ and $K_h = 5$.

Table 2 lists the 10 most frequent words over 20 observed labels and 5 hidden topics. This result is similar to that given by DiscLDA [12], where words that occurred in the labeled rows are significantly cross-referenced with their corresponding labels, and words in the hidden topics show no obvious preference to any class.

5.2 Classification performance

For single-label corpus, SL-LDA assigns each testing document d a label y_d^* by:

$$y_d^* = \underset{k=1,2,\cdots,K_t}{\arg\max}\ (\theta_{d,k}) \tag{11}$$

Following previous studies [12, 23], we evaluated SL-LDA in terms of binary- and multi-class document classification. Several existing supervised topic models were chosen as performance baselines, i.e., LDA-SVM, sLDA, DiscLDA, and MedLDA. Average scores were obtained from 20 runs, and pairwise t-tests between SL-LDA and the baselines were conducted at the 5 % significance level. As in [22], an indicator •/◦ is used to denote whether SL-LDA was found to be statistically superior/inferior to the compared algorithm.

5.2.1 Binary classification

The binary-class ($K_t = 2$) classification experiments were performed on two subgroups of Newsgroups, i.e., alt.atheism and talk.religion.misc, following the design described in [12, 23].

For the LDA-SVM, we used the Gibbs-EM algorithm to fit the upstream LDA model, and employed the celebrated LibSVM[4] as the downstream classifier. We used a radial basis function as the kernel, and optimized its parameters via the grid search method. The public codes of sLDA[5] and MedLDA[6], were employed for dependable results. Their parameters were determined according to the discussions in the corresponding publications. The DiscLDA results were taken from the original paper, as we could not obtain its primary implementation.

Table 3 lists the results for different topic numbers K. We can see that SL-LDA attains a competitive level of performance, and is statistically superior to other models in most cases. Our model obtains better scores for relatively small K, and achieves the highest score of 0.819 when K=15. As

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

[3] http://mallet.cs.umass.edu

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

[5] http://www.cs.princeton.edu/~blei/topicmodeling.html

[6] http://www.ml-thu.net/~jun/software.html

Table 2 The most frequent words over artificial assigned labels and hidden labels

Topic	The most frequent words
Alt.atheism	atheism; alt; sgi; god; writes; livesey; article; morality; wpd; solntze
Comp.graphics	graphics; comp; de; image; fi; bit; berlin; alt; files; computers
Comp.os.ms.windows.misc	windows; comp; misc; win; file; program; oracle; files; unix; system
Comp.sys.ibmpc.hardware	ibm; sys; pc; hardware; drive; comp; austin; port; card; software
Comp.sys.mac.hardware	mac; comp; sys; hardware; apple; drive; problem; monitor; system; mhz
Comp.windows.x	windows; comp; window; server; mit; motif; text; sun; code; program
Misc.forsale	forsale; misc; sale; du; computers; nyx; usenet; rec; distribution; sender
Rec.autos	uiuc; rec; autos; car; cars; ux; writes; illinois; reston; usenet
Rec.motorcycles	motorcycles; rec; bike; article; org; writes; sender; mitre; ride; rider
Rec.sport.baseball	baseball; rec; sport; year; game; cubs; writes; team; rochester; hit
Rec.sport.hockey	hockey; game; sport; rec; team; play; year; players; games; season
Sci.crypt	sci; crypt; key; clipper; security; encryption; alt; org; chip; privacy
Sci.electronics	electronics; sci; audio; input; pin; signal; circuit; work; copy; data
Sci.med	sci; med; food; disease; medical; people; alt; energy; writes; misc
Sci.space	space; sci; launch; shuttle; digex; henry; mission; orbit; access; article
Soc.religion.christian	rutgers; christian; god; religion; church; soc; geneva; jesus; igor; aramis
Talk.politics.guns	guns; gun; politics; talk; stratus; people; government; alt; writes; weapons
Talk.politics.mideast	soc; culture; talk; politics; mideast; turkish; israel; jewish; israeli; greek
Talk.politics.misc	misc; politics; talk; alt; writes; people; article; clinton; government; legal
Talk.religion.misc	talk; religion; misc; alt; writes; abortion; article; apple; god; frank
Hidden topic 1	washington; power; air; speed; test; heat; boeing; article; clock; plastic
Hidden topic 2	sun; send; list; request; requests; group; community; ebay; groups; tools
Hidden topic 3	state; ohio; mps; zaphod; sei; cc; cis; club; magnesium; apr
Hidden topic 4	purdue; option; station; redesign; ecn; human; colostate; capability; committee; freedom
Hidden topic 5	att; arizona; cb; princeton; uchicago; uchinews; linac; writes; uwm; convenient

K becomes larger, SL-LDA performs sightly worse than the state-of-the-art MedLDA, e.g., by 0.09 for K=25 and 0.02 for K=35. We believe this is because larger values of K bring too many hidden topics for binary-class settings, resulting in reduced performance. In Summary, our model can provide higher performance with fewer topics, i.e., with less computational expense. This characteristic is quite meaningful in practice.

5.2.2 Mutli-class classification

The multi-class classification experiments considered the full Newsgroups collection. We compared SL-LDA with LDA-SVM, sLDA, and MedLDA. All three benchmark models were set up as for the binary-class experiment.

The experimental results are given in Table 4. Clearly, SL-LDA achieves the top scores for all K, and is also

Table 3 The performance (averaged score± standard deviation) for binary-class cases; •/○ means whether SL-LDA is statistically superior/inferior to the compared algorithm

Topics	SL-LDA	LDA-SVM	sLDA	MedLDA	DiscLDA
5	0.786±0.013	0.672±0.035•	0.621±0.039•	0.762±0.018•	**0.800**± * **
10	**0.819±0.004**	0.692±0.016•	0.636±0.027•	0.782±0.009•	0.800± * **
15	**0.807±0.006**	0.682±0.012•	0.642±0.021•	0.792±0.005	0.800± * **
20	0.792±0.005	0.692±0.008	0.667±0.016•	**0.802±0.008**	0.800± * **
25	0.795±0.003	0.696±0.009•	0.689±0.011•	**0.804±0.005**	0.800± * **
30	0.797±0.004	0.673±0.007•	0.703±0.008•	**0.809±0.004**	0.800± * **
35	0.799±0.002	0.652±0.007•	0.682±0.006•	**0.801±0.005•**	0.800±* * *

Bold entries denote the best scores

Table 4 The performance (averaged score± standard deviation) for multi-class cases; •/○ means whether SL-LDA is statistically superior/inferior to the compared algorithm

Topics	SL-LDA	LDA-SVM	sLDA	MedLDA
30	**0.806±0.008**	0.609±0.028•	0.506±0.016•	0.796±0.012•
40	**0.816±0.004**	0.626±0.034•	0.472±0.029•	0.797±0.011•
50	**0.827±0.007**	0.663±0.016•	0.498±0.011•	0.793±0.014•
60	**0.824±0.011**	0.654±0.032•	0.536±0.009○	0.806±0.013
70	**0.828±0.008**	0.691±0.009	0.607±0.018•	0.802±0.009
80	**0.840±0.009**	0.686±0.043•	0.609±0.018•	0.809±0.009
90	**0.836±0.005**	0.678±0.025•	0.589±0.014•	0.801±0.008•
100	**0.839±0.016**	0.679±0.018	0.595±0.026•	0.822±0.015○
110	**0.838±0.009**	0.702±0.014•	0.611±0.017•	0.809±0.012

Bold entries denote the best scores

statistically superior to the other models in most cases. Compared with LDA-SVM and sLDA, SL-LDA yields significant improvements, scoring around 0.2 higher than LDA-SVM and 0.3 higher than sLDA. More importantly, SL-LDA outperforms the state-of-the-art MedLDA by about 0.02.

5.2.3 Study on parameter θ^*

In this section, we examine the label threshold θ^*. Figure 1 illustrates the binary-class results with different values of θ^*. We can clearly observe that the performance at values of 0.5 and 0.6 is better in than the other cases. The multi-class cases (Fig. 2) exhibit a similar trend, with 0.5 and 0.6 dominating the performance. These results conform to the discussions in Section 2.2.

5.3 Running time

We now examine the time efficiency of SL-LDA on a 3.1GHz Intel Core i5 2400 CPU. To ensure a fair comparison, two public models, i.e., sLDA and MedLDA, were used as baselines.

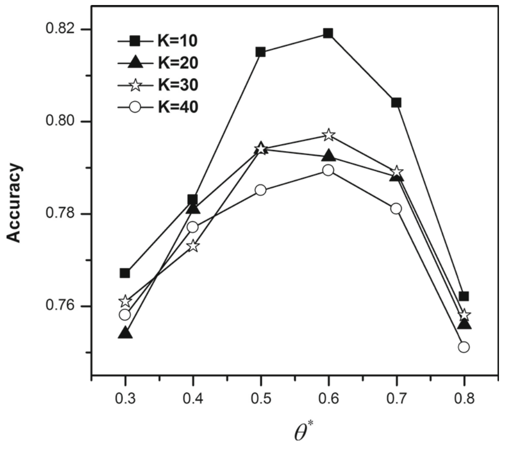

Fig. 1 The evaluation of parameter θ^* on binary-class case

Figure 3 shows that SL-LDA is more efficient than sLDA. As reported in [24], we found that sLDA is quite time-consuming. This is mainly because the normalization factor strongly couples the topic assignments of all the words. SL-LDA is also faster than MedLDA, which needs to solve an extra dual problem during training.

6 Evaluation on multi-label setting

In this section, we evaluate the performance of SL-LDA for multi-label ($K_d > 1$) document classification. The parameter θ^* is tuned using values in the set $\{0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8\}$, and the other parameters are set as for the single-label case.

6.1 Metric

The multi-label classification problem requires more metrics than the single-label case. In this experiment, we employ several popular metrics [18] to evaluate SL-LDA.

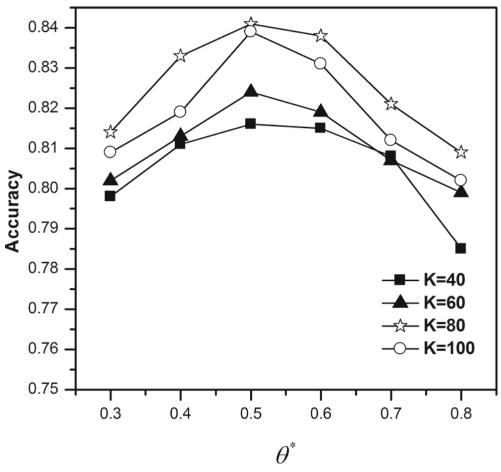

Fig. 2 The evaluation of parameter θ^* on multi-class corpus

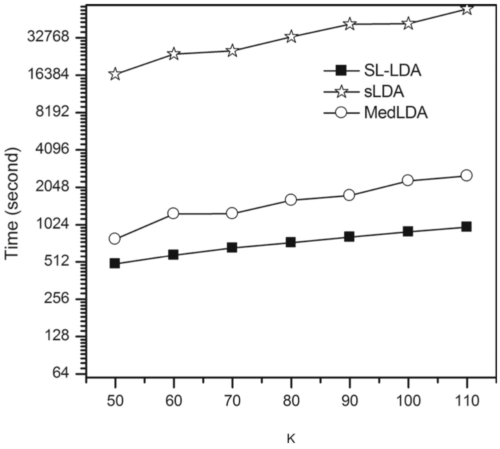

Fig. 3 The training time (seconds in log2-scale) in terms of different number of topics for multi-class classification

Assume that the test corpus W' consists of D' documents. For each document d, y_d and y_d^* denote the true and estimated label set, respectively, where $y_d, y_d^* \subseteq \{1, 2, \cdots, K_t\}$.

6.1.1 Rank-based metric

The estimated rank of label k for document d is denoted by $r_d(k)$. We introduce three rank-based metrics, for each of which smaller values imply better classification.

Ranking loss This measures the number of times that irrelevant labels are ranked higher than relevant labels:

$$Rnk-Loss = \frac{1}{D'}\sum_{d=1}^{D'} \times \frac{1}{|y_d|\,|\overline{y_d}|}\left|\left\{(k_i,k_j) : r_d(k_i) > r_d(k_j), (k_i,k_j) \in y_d \times \overline{y_d}\right\}\right| \tag{12}$$

where $\overline{y_d}$ is the complement of y_d .

One error This measures how many times the top-ranked label is not in the true label set:

$$One-Err = \frac{1}{D'}\sum_{d=1}^{D'} \delta\left(\underset{k=1,2,\cdots,K_t}{argmin}\; r_d(k)\right) \tag{13}$$

where

$$\delta(k) = \begin{cases} 1 & k \notin y_d \\ 0 & otherwise \end{cases}$$

Margin Measures the difference in ranking between the top-ranked irrelevant label and the lowest-ranked relevant label:

$$Margin = \frac{1}{D'}\sum_{d=1}^{D'}\left|\underset{k\in\overline{y_d}}{argmin}\, r_d(k_k) - \underset{k\in y_d}{argmax}\, r_d(k_k)\right| \tag{14}$$

6.1.2 Binary prediction metric

The two binary prediction metrics used in our experiments are the Macro-F1 and Micro-F1 scores. Larger Macro-F1 and Micro-F1 scores denote better performance.

We define the Recall, Precision and F1-score for a document d as follows:

$$Recall\,(d) = \frac{|y_d \cap y_d^*|}{|y_d|}$$

$$Precision\,(d) = \frac{|y_d \cap y_d^*|}{|y_d^*|}$$

$$F1-score\,(d) = \frac{2 \times Recall\,(d) \times Precision\,(d)}{Recall\,(d) + Precision\,(d)} \tag{15}$$

After computing the F1-scores for all the test documents, the Macro-F1 metric is obtained by averaging all of the individual F1-scores. The Micro-F1 considers the full testing corpus as a large document. It can be directly computed using (15).

6.2 Datasets

Yahoo! Arts and Health [7] These two corpora are from the Yahoo! Collection. Arts consists of 7,741 documents and 19 unique labels, and Health contains 9,109 documents and 14 unique labels. The *cardinality* of both datasets is relatively small, i.e., 1.7 (Arts) and 1.6 (Health), and about 55% of the documents have only a single label.

Following the preprocessing steps in [10], we randomly sampled 1,000 documents from each dataset, ensuring that each label appeared at least once, to form the training data. The remaining documents were used as the test data. This process was repeated 5 times to give 5 available training/test splits.

RCV1-v2 [13] The RCV1-v2 dataset is another popular benchmark for multi-label document classification. It consists of over 800,000 documents with 103 labels. In our experiment, we used the original training set from the *LYRL2004* split [13], which contains 23,149 documents assigned by 101 labels. We selected 30,000 documents at random from the *LYRL2004* test split to form the test data. This procedure was repeated 10times, giving 10 training/test splits. The *cardinality* of RCV1-v2 is about 3.1, larger than that of the Yahoo! subdirectories.

6.3 Classification performance

Several existing approaches were compared with SL-LDA, including L-LDA [15], Pr-LDA, Dep-LDA [17], and Tuned-SVMs (T-SVMs) [13]. We implemented the in-house codes

[7] http://www.kecl.ntt.co.jp/as/members/ueda/yahoo.tar.gz

Table 5 The Experiment results (averaged score± standard deviation) on Ranking Loss (top section), One Error (middle section) and Margin (bottom section); ●/○ means whether SL-LDA is statistically superior/inferior to the compared algorithm

	SL-LDA	L-LDA	Prior-LDA	Dep.-LDA	T-SVMs
Arts	**0.145±0.002**	0.194±0.008●	0.149±0.005●	0.146±0.004●	0.159±0.005●
Health	**0.073±0.003**	0.104±0.009●	0.092±0.008●	0.075±0.007	0.102±0.012●
RCV1-v2	**0.012±0.001**	0.053±0.004●	0.031±0.002●	0.013±0.001	0.013±0.001
Arts	**0.445±0.006**	0.482±0.013●	0.474±0.011●	0.459±0.009●	0.458±0.004○
Health	0.244±0.003	0.336±0.018●	0.332±0.009●	**0.241±0.003**○	0.246±0.004●
RCV1-v2	**0.059±0.009**	0.196±0.058●	0.129±0.023●	0.069±0.013	0.061±0.019●
Arts	**3.52±0.013**	5.14±0.011●	3.73±0.031●	3.88±0.022●	4.17±0.009
Health	**1.42±0.007**	2.11±0.018●	1.91±0.026●	1.49±0.015●	2.09±0.009●
RCV1-v2	2.98±0.026	13.49±0.098●	8.92±0.057●	**2.86±0.028**	2.89±0.042●

Bold entries denote the best scores

for these three supervised LDA models. All parameters were set according to the discussions in the original papers. We implemented T-SVMs using LibSVM and the parameters in [13]. Each approach was executed 20 times on each training/test split (i.e., a total of 100 times for Arts and Health and 200 times for RCV1-v2), and pairwise t-tests between SL-LDA and the baselines were conducted at the 5 % significance level.

In the early experiments, we found the performance to be stable while the number of topics K was slightly larger than the true number of labels, but to decrease for significantly larger values of K. We argue that this is reasonable, because, intuitively, a large K will exaggerate the effect of hidden topics. In this section, we report the results for the two Yahoo! datasets and the RCV1-v2 dataset with $K = 100$ and $K = 240$, respectively.

6.3.1 Rank-based performance

SL-LDA outputs the distribution of topics for each unlabeled document. Thus, we can directly rank the existing label k for document d , i.e., $r_d(k)$. The experiment results are given in Table 5: our model performs well in terms of both numerical results and statistics.

Among the supervised LDA models, SL-LDA performs much better than L-LDA and Pr-LDA, and achieves competitive performance with the state-of-the-art Dep-LDA. The difference between L-LDA and SL-LDA is significant. The difference between Pr-LDA and SL-LDA is also large, except for the Yahoo! Arts dataset. We believe this is because of the small size and low label density of this dataset. SL-LDA also outperformed the state-of-the-art Dep-LDA on 5/6 scores across the Yahoo! datasets and on 2/3 evaluation metrics across the larger RCV1-v2 dataset. Although SL-LDA is simpler than Dep-LDA, it attains better performance in terms of rank-based metrics.

SL-LDA also achieved higher scores than T-SVMs for 8/9 metrics across all three datasets. These results demonstrate that SL-LDA is competitive with the state-of-the-art discriminative method.

6.3.2 Binary prediction performance

To compute the binary prediction metrics, we must transform the label ranking of each test document d into its estimated label set y_d^*, which consists of the top N ranked documents. Here, we set N equal to the *cardinality* of each dataset.

The binary prediction results are listed in Table 6. We can see that SL-LDA is statistically superior to the other methods, and attains better performance. In most cases, SL-LDA outperforms the other supervised LDA models in terms of both Macro-F1 and Micro-F1 across all three datasets. For the simpler models, SL-LDA scores about 0.09-0.11 higher

Table 6 The Experiment results (averaged score± standard deviation) on Macro-F1 (top section) and Micro-F1 (bottom section); ●/○ means whether SL-LDA is statistically superior/inferior to the compared algorithm

	SL-LDA	L-LDA	Prior-LDA	Dep.-LDA	T-SVMs
Arts	**0.518±0.008**	0.428±0.013●	0.472±0.008●	0.499±0.009●	0.499±0.009
Health	**0.712±0.009**	0.614±0.033●	0.635±0.017●	0.711±0.021●	0.702±0.007
RCV1-v2	**0.831±0.005**	0.539±0.018●	0.603±0.013●	0.828±0.008	0.829±0.0011●
Arts	0.478±0.004	0.384±0.017●	0.448±0.011●	0.473±0.007●	**0.485±0.004**
Health	**0.695±0.006**	0.583±0.006	0.614±0.008●	0.683±0.008●	0.668±0.011●
RCV1-v2	0.792±0.007	0.506±0.023●	0.562±0.015●	0.786±0.012●	**0.801±0.007**

Bold entries denote the best scores

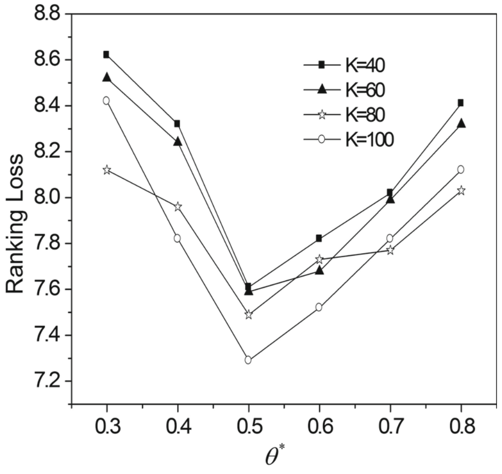

Fig. 4 The evaluation of parameter θ^* on Yahoo! Arts collection

than L-LDA and about 0.04-0.08 higher than Pr-LDA on the Yahoo! datasets, and achieves greater improvements across the larger RCV1-v2 dataset. In particular, SL-LDA outscores the state-of-the-art Dep-LDA by about 0.005-0.02. Regarding T-SVMs, we can see that the proposed model scores slightly lower on the RCV1-v2 dataset (similar results have been reported in [13]). However, our method outperforms T-SVMs on 3/4 metrics across the smaller Yahoo! datasets(an improvement of around 0.02 in Macro-F1 across Yahoo! Arts and 0.03 in Micro-F1 across Yahoo! Health). These results indicate that SL-LDA is competitive with state-of-the-art approaches.

6.4 Study on parameter θ^*

We now study the effect of parameter θ^* in the multi-label setting. Two datasets are used: 1) Yahoo! Arts collection (which has a small *cardinality* of 1.7); and 2) RCV1-v2 collection (which has a larger *cardinality* of about 3.1). As different metrics exhibit similar trends, we only use the ranking loss to compare different values of θ^* across the two datasets.

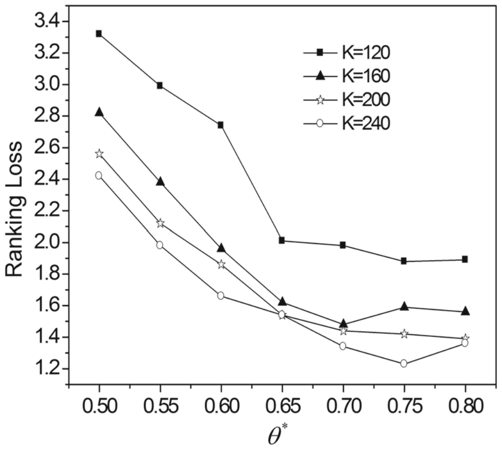

Fig. 5 The evaluation of parameter θ^* on RCV1-v2 collection

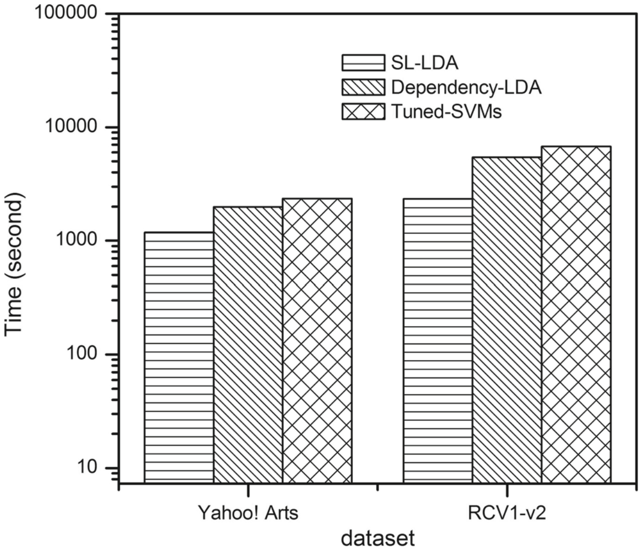

Fig. 6 The training time (seconds in log10-scale) for multi-label classification

The experimental results for the Yahoo! Arts collection are shown in Fig. 4. It is clear that higher performance is obtained when $\theta^* = 0.5$ or 0.6. Larger values tend to degrade the performance. The trend for the Yahoo! Arts collection is very similar to that of the single-label corpus in Section 5. This is because Yahoo! Arts contains fewer labels, and about half the documents are tagged with only a single label.

The results for the RCV1-v2 collection (Fig. 5) are somewhat different, with larger values of θ^* tending to improve performance. The peak score is obtained for a value of $\theta^* = 0.75$. Because documents in RCV1-v2 contain an average of around three labels, a higher weighted proportion of pre-assigned labels are needed to describe a labeled document.

Clearly, the experimental results also conform to the discussions in Section 2.2. However, in addition to the multi-label settings, we suggest that a small/large value of θ^* should be applied for datasets with small/large *cardinality*.

6.5 Running time

Finally, we compared the time efficiency of SL-LDA with that of two state-of-the-art approaches, i.e., Dep-LDA and T-SVM, for multi-label classification. Pr-LDA was not chosen, because it was deemed to be a special case of Dep-LDA.

Figure 6 shows the running time of different approaches for the Yahoo! Arts[8] and RCV1-v2 collections. Clearly, SL-LDA[9] is more efficient than Dep-LDA and T-SVMs. For

[8] Results for the Yahoo! Health collection are very similar.

[9] We set $K = 100$ for Yahoo! Arts and $K = 240$ for RCV1-v2.

the smaller Yahoo! Arts dataset, SL-LDA is slightly faster than Dep-LDA, but the efficiency gap widens for the RCV1-v2 dataset. This is because Dep-LDA adds a label layer to the model, and its computational complexity is sensitive to the number of labels. The slowness of T-SVMs is mainly because of the parameter optimization process. In contrast, SL-LDA has few parameters, and is therefore faster than T-SVMs in this step. Following our discussion on the choice of $\theta*$, we can quickly determine an appropriate value. Thus, we argue that SL-LDA is more efficient in practice than Dep-LDA and T-SVMs.

7 Conclusion

In this paper, we have suggested a novel supervised model for document classification, including both single-label and multi-label settings. Based on the L-LDA model, SL-LDA relaxes the restriction of label sampling, and extends the topics concept to capture lost labels and common semantics. These modifications significantly improve the classification performance. We developed a Gibbs-EM algorithm to estimate and infer our model. A series of evaluation experiments were conducted, and the results show that: (1) in single-label cases, SL-LDA outperforms LDA-SVM, sLDA, DiscLDA, and the state-of-the-art MedLDA in most instances; (2) SL-LDA significantly outperforms L-LDA and Pr-LDA, and, more importantly, is competitive with the state-of-the-art Dep-LDA and SVMs.

In the future, we intend to: (1) develop online approaches for large-scale multi-label corpora; (2) investigate collections that contain many labels; (3) apply our model to some other applications, e.g., summarization and filtering.

Acknowledgments This work was supported by National Nature Science Foundation of China (NSFC) under the Grant No. 61170092, 61133011, and 61103091.

References

1. Ali D, Faqir M (2012) Group topic modeling for academic knowledge discovery. Appl Intell 36(4):870–886
2. Andrieu C, Freitas ND, Doucet A, Jordan MI (2003) An introduction to MCMC for machine learning. Mach Learn 50(1):5–43
3. Blei DM, Lafferty JD (2007) A correlated topic model fo science. Ann Appl Stat 1(1):17–35
4. Blei DM, McAuliffe JD (2007) Supervised topic models. In: Neural information processing systems
5. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
6. Fei-Fei L, Perona P (2005) A Bayesian hierarchical model for learning natural scene categories. In: IEEE computer society conference on computer vision and pattern recognition, vol 2, pp 524–531
7. Heinrich G. (2005) Parameter estimation for text analysis. http://www.arbylon.net/publications/textest
8. Hofmann T (1999) Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp 50–57
9. Jaegul C, Changhyun L, Chandan KR, Park H (2013) Utopian: user-driven topic modeling based on interactive nonnegative matrix factorization. IEEE Trans Vis Comput Graph 19(12):1992–2001
10. Ji S, Tang L, Yu S, Ye J (2008) Extracting shared subspace for multi-label classification. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, pp 381–389
11. Kim D, Kim S, Oh A (2012) Dirichlet process with mixed random measures: a nonparametric topic model for labeled data. In: 29th International conference on machine learning, pp 727–734
12. Lacoste-Julien S, Sha F, Jordan MI (2009) Disclda: discriminative learning for dimensionality reduction and classification. In: Neural information processing systems, pp 897–904
13. Lewis DD, andTony G, Rose YY, Li F (2004) Rcv1: a new benchmark collection for text categorization research. J Mach Learn Res 5:361–397
14. Quelhas P, Monay F, Odobez JM, Gatica-Perez D, Tuytelaars T, Van Gool L (2005) Modeling scenes with local descriptors and latent aspects. Comput Vis IEEE Int Conf 1:883–890
15. Ramage D, Hall D, Nallapati R, Manning CD (2009) Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: Conference on empirical methods in natural language processing, pp 248–256. Association for Computational Linguistics
16. Ramage D, Manning CD, Dumais S (2011) Partially labeled topic models for interpretable text mining. In: ACM SIGKDD international conference on knowledge discovery and data mining, pp 457–465
17. Rubin TN, Chambers A, Smyth P, Steyvers M (2012) Statistical topic models for multi-label document classification. Mach Learn 88(1–2):157–208
18. Sebastiani F (2002) Machine learning in automated text categorization. ACM Comput Surv (CSUR) 34(1):1–47
19. Wallach H (2006) Topic modeling: beyond bag-of-words. In: Proceedings of the 23rd international conference on Machine learning, pp 977–984. ACM
20. Xie P, Xing EP (2013) Integrating document clustering and topic modeling. In: Proceedings of the 20th conference on uncertainty in artificial intelligence, pp 694–703
21. Xu Y, Guo R (2014) An inproved nu-twin support vector machine. Appl Intell 41(1):42–54
22. Zhang ML, Zhang K (2010) Multi-label learning by exploiting label dependency. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, pp 999–1008
23. Zhu J, Ahmed A, Xing E (2009) Medlda: maximum margin supervised topic models for regression and classification. In: Proceedings of the 26th annual international conference on machine learning, pp 1257–1264. ACM
24. Zhu J, Ahmed A, Xing E. (2012) Medlda: maximum margin supervised topic models

Ximing Li received the M.S. degree in computer science from Jilin University, China, in 2011. Currently he is a Ph.D. candidate in the College of Computer Science and Technology at Jilin University. His main research interests include topic modeling and multi-label learning.

Jihong Ouyang is a professor at the College of Computer Science and Technology, Jilin University of China. She received her Ph.D degree in Jilin university in 2005. Her main research interests include artificial intelligence and machine learning, more specifically in spatial reasoning, multi-label learning, topic modeling, and online learning.

Xiaotang Zhou received the M.S. degree in computer science from Jilin University, China, in 2013. Currently he is a Ph.D. candidate in the College of Computer Science and Technology at Jilin University. His main research interests include data mining and topic modeling.

You Lu received the B.S. degree in computer science from Jilin University, China, in 2012. Currently he is an M.S. candidate in the College of Computer Science and Technology at Jilin University. His main research interests include topic modeling and variational inference.

Yanhui Liu received the B.S. degree in computer science from Jilin University, China, in 2012. Currently he is an M.S. candidate in the College of Computer Science and Technology at Jilin University. His main research interests include topic modeling and semantics analysis.