Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Supervised topic models with weighted words: multi-label document classification*

Yue-peng ZOU[1,2], Ji-hong OUYANG[†1,2], Xi-ming LI[†‡1,2]

[1]College of Computer Science and Technology, Jilin University, Changchun 130012, China
[2]MOE Key Laboratory of Symbolic Computation and Knowledge Engineering, Jilin University, Changchun 130012, China
[†]E-mail: ouyj@jlu.edu.cn; liximing86@gmail.com

Received Oct. 26, 2016; Revision accepted Jan. 3, 2017; Crosschecked Apr. 3, 2018

Abstract: Supervised topic modeling algorithms have been successfully applied to multi-label document classification tasks. Representative models include labeled latent Dirichlet allocation (L-LDA) and dependency-LDA. However, these models neglect the class frequency information of words (i.e., the number of classes where a word has occurred in the training data), which is significant for classification. To address this, we propose a method, namely the class frequency weight (CF-weight), to weight words by considering the class frequency knowledge. This CF-weight is based on the intuition that a word with higher (lower) class frequency will be less (more) discriminative. In this study, the CF-weight is used to improve L-LDA and dependency-LDA. A number of experiments have been conducted on real-world multi-label datasets. Experimental results demonstrate that CF-weight based algorithms are competitive with the existing supervised topic models.

Key words: Supervised topic model; Multi-label classification; Class frequency; Labeled latent Dirichlet allocation (L-LDA); Dependency-LDA
https://doi.org/10.1631/FITEE.1601668 **CLC number:** TP391

1 Introduction

Latent Dirichlet allocation (LDA) (Blei et al., 2003) is a probabilistic Bayesian model (Ghahramani, 2001) used to process discrete data, e.g., text document collections. During the past decade, this model has been studied extensively and successfully applied to various document-related tasks, such as classification, clustering, and summarization. Because the original LDA model is unsupervised, the development of supervised modifications for classification is one of the most active topics in topic modeling research. Nowadays, the representative supervised topic models include: supervised LDA (Blei and McAuliffe, 2007), discriminative LDA (Lacoste-Julien et al., 2008), and maximum entropy discrimination LDA (Zhu et al., 2012) for single-label classification; labeled LDA (L-LDA) (Ramage et al., 2009), partially labeled LDA (Ramage et al., 2011), Dirichlet process with mixed random measures (Kim et al., 2012), dependency-LDA (Rubin et al., 2012), and dependency frequency LDA (DF-LDA) (Li et al., 2015b) for multi-label classification.

To our knowledge, L-LDA is the first supervised LDA model for multi-label document classification. It incorporates the supervision, i.e., the class label, into the LDA model by first defining a 1:1 correspondence between labels and topics, and then constraining each document to its pre-assigned label set. Dependency-LDA and DF-LDA further extend L-LDA by considering the label frequency and label dependency. These models have achieved competitive empirical results against popular discriminative algorithms. However, the class frequency information for words—the

‡ Corresponding author
* Project supported by the National Natural Science Foundation of China (No. 61602204)
ORCID: Xi-ming LI, http://orcid.org/0000-0001-8190-5087

number of classes where a word has occurred in the training data—is neglected. That is, these models are concerned mainly about the intra-label word weight, but neglect the inter-label word weight, resulting in a lack of feature selection.

To address the problem mentioned above, we attempt to add a word (i.e., feature) weight assignment step into the supervised topic models. In this study, we weight words by the class frequency, which is a more discriminative knowledge classification (Guan et al., 2009; Li et al., 2015a). Note that if a word occurs only in a few labels (i.e., low-class frequency), it should be discriminative for these labels. In this sense, we propose a word weighting method, namely class frequency weight (CF-weight), which provides low (high) class frequency words with large (small) weights. In this study, we use the CF-weight to improve L-LDA and dependency-LDA models for multi-label document classification. To evaluate the performance of the proposed models, a number of experiments have been conducted on real-world multi-label datasets. Experimental results demonstrate that the supervised topic models with the CF-weight can improve the multi-label classification performance. Some important notations are summarized in Table 1.

Table 1 Notation description

Notation	Description
D	Number of documents
V	Number of words
C	Number of labels
T	Number of topics in dependency-LDA
$\boldsymbol{z}$	Label assignments
$(\alpha, \boldsymbol{\theta})$	Document-label Dirichlet-multinomial pair in L-LDA
$(\beta, \boldsymbol{\phi})$	Label-word Dirichlet-multinomial pair
$\boldsymbol{z}'$	Topic assignments in dependency-LDA
$\boldsymbol{c}'$	Label assignments from $\boldsymbol{\phi}$ in dependency-LDA
$(\alpha', \boldsymbol{\theta}')$	Document-topic Dirichlet-multinomial pair in dependency-LDA
$(\beta', \boldsymbol{\phi}')$	Topic-label Dirichlet-multinomial pair in dependency-LDA
$(\boldsymbol{\alpha}_d, \boldsymbol{\theta})$	Document-label Dirichlet-multinomial pair in dependency-LDA

2 Background

2.1 Labeled latent Dirichlet allocation (L-LDA)

L-LDA (Ramage et al., 2009) is an extension of the unsupervised LDA model. To incorporate the supervision, it applies a 1:1 correspondence between topics and labels, and assumes that each document d is restricted to a description by a multinomial distribution $\boldsymbol{\theta}_d$ over labels included in its label set y_d, where each label c is a multinomial distribution $\boldsymbol{\phi}_c$ over words. The generative process of L-LDA is shown in Algorithm 1.

During model training, the goal of L-LDA is to estimate all C label-word distributions $\boldsymbol{\phi}$. During the testing, L-LDA predicts test documents by estimating the document-label distributions $\boldsymbol{\theta}$. Because the label sets of the test documents are unknown, L-LDA reduces to the unsupervised LDA model during the testing.

2.2 Dependency-LDA

Recently, dependency-LDA (Rubin et al., 2012) further considers the label frequency and label dependency observed in the training data by using an asymmetric document-label Dirichlet prior. To capture the label frequency, dependency-LDA introduces a corpus-wide distribution $\boldsymbol{\varphi}$ over the labels (i.e., a power-law distribution of label frequency), drawn from a Dirichlet prior β'. For each document d, it independently draws M_d label tokens from distribution $\boldsymbol{\varphi}$, and then constructs its asymmetric document-label Dirichlet prior $\boldsymbol{\alpha}_d$ as follows:

$$\boldsymbol{\alpha}_d = \left[\eta_d \frac{N_{d,1}}{M_d} + \alpha_d,\ \eta_d \frac{N_{d,2}}{M_d} + \alpha_d,\ \cdots,\ \eta_d \frac{N_{d,C}}{M_d} + \alpha_d \right], \tag{1}$$

where $N_{d,i}$ is the number of times that label i has been sampled from $\boldsymbol{\varphi}$, η_d is the scale coefficient, and α_d is the smoothing parameter. Given the prior $\boldsymbol{\alpha}_d$, the following word generation process is the same as that in L-LDA.

Algorithm 1 Generative process of L-LDA

for each label $c \in \{1, 2, \ldots, C\}$ **do**
  Sample a distribution over words $\boldsymbol{\phi}_c \sim \text{Dirichlet}(\beta)$
end for
for each document $d \in \{1, 2, \ldots, D\}$ **do**
  Sample a distribution over labels in its label set y_d: $\boldsymbol{\theta}_d \sim \text{Dirichlet}(\alpha)$
  for each word $w_{d,n}$ of N_d words **do**
    Sample a label $z_{d,n} \sim \text{Multinomial}(\boldsymbol{\theta}_d)$
    Sample a word $w_{d,n} \sim \text{Multinomial}(\boldsymbol{\phi}_{z_{d,n}})$
  end for
end for

To further capture the label dependencies, dependency-LDA assumes that there exist T corpus-wide distributions $\boldsymbol{\varphi}_t$ of label frequency, referred to as a 'topic'. To construct the asymmetric Dirichlet prior $\boldsymbol{\alpha}_d$ of each document d, it first samples a document-topic distribution $\boldsymbol{\theta}_d$ from a Dirichlet prior α', and then repeats the following process M_d times for M_d label tokens: (1) sampling a topic z' from $\boldsymbol{\theta}_d$; (2) sampling a label token c' from topic $\boldsymbol{\varphi}_{z'}$; (3) finally computing $\boldsymbol{\alpha}_d$ using Eq. (1). The generative process of dependency-LDA is shown in Algorithm 2.

During model training, the goal of dependency-LDA is to estimate all T topic-label distributions $\boldsymbol{\varphi}$ and all C label-word distributions $\boldsymbol{\phi}$. During the testing, dependency-LDA also predicts test documents by estimating the document-label distributions $\boldsymbol{\theta}$.

Algorithm 2 Generative process of dependency-LDA

for each topic $t \in \{1, 2, \ldots, T\}$ **do**
  Sample a distribution over words $\boldsymbol{\varphi}_t \sim \text{Dirichlet}(\beta')$
end for
for each label $c \in \{1, 2, \ldots, C\}$ **do**
  Sample a distribution over words $\boldsymbol{\phi}_c \sim \text{Dirichlet}(\beta)$
end for
for each document $d \in \{1, 2, \ldots, D\}$ **do**
  Sample a distribution over topics $\boldsymbol{\theta}_d \sim \text{Dirichlet}(\alpha')$
  for each label $i \in \{1, 2, \ldots, M_d\}$ **do**
    Sample a topic $z'_{d,i} \sim \text{Multinomial}(\boldsymbol{\theta}_d')$
    Sample a label in y_d: $c'_{d,i} \sim \text{Multinomial}(\boldsymbol{\varphi}_{z'_{d,i}})$
  end for
  Compute the asymmetric Dirichlet prior $\boldsymbol{\alpha}_d$ using Eq. (1)
  Sample a distribution over labels in y_d: $\boldsymbol{\theta}_d \sim \text{Dirichlet}(\alpha)$
  for each word $w_{d,n}$ of N_d words **do**
    Sample a label $z_{d,n} \sim \text{Multinomial}(\boldsymbol{\theta}_d)$
    Sample a word $w_{d,n} \sim \text{Multinomial}(\boldsymbol{\phi}_{z_{d,n}})$
  end for
end for

3 Weighting the words

Supervised topic models have been successfully applied to multi-label document classification (Ramage et al., 2009; Rubin et al., 2012; Li et al., 2015b). However, they neglect the class frequency information for words, which is quite significant for classification. Taking L-LDA as an example, this problem can be described as follows: Given an observed collection, L-LDA will estimate all C label-word distributions $\boldsymbol{\phi}$, which are independent of each other. If word v occurs frequently in label c, its value for label-word probability $\varphi_{c,v}$ may be large. That is, word v will be significant in predicting whether a test document belongs to label c. However, if word v also occurs frequently in most of the other labels (i.e., large class frequency), it is less discriminative. This leads to a conflict, and may reduce the accuracy of classification.

To address the problem mentioned above, we investigate weighting words in supervised topic models. For each word v, we compute its weight value based on the class frequency (CF-weight) as

$$\text{CFW}_v = \eta \log\left(\frac{C}{\text{CF}_v}\right) + \gamma, \tag{2}$$

where CF_v is the class frequency of word v, i.e., the number of labels where word v has occurred in the training data, η is the scale coefficient, and γ is the smoothing factor. Two parameters, η and γ, are used to tune the influence of the CF-weight.

A word with a higher (lower) class frequency will correspond to a smaller (larger) CF-weight (Eq. (2)). This is based on the thinking that a word with a higher/lower class frequency will be less/more discriminative for classification. For supervised topic models, CF-weight can be considered to be the soft count of words based on the class frequency. In this study, we use the CF-weight to improve L-LDA and dependency-LDA. An example of L-LDA is shown in Fig. 1.

3.1 Weighting the words for L-LDA

CF-weight is used for L-LDA, and then a weighted L-LDA (WL-LDA) is proposed. In WL-LDA, each word token is weighted by its corresponding CF-weight.

3.1.1 Model training

The goal of model training for WL-LDA is to estimate all C label-word distributions $\boldsymbol{\phi}$. Given an observed data W, this is achieved by approximating the posterior distribution of latent variables $p(\boldsymbol{\phi}, \theta, \boldsymbol{z}|W, \alpha, \boldsymbol{\beta})$. Both variational inference (VI) and the Gibbs sampler are used in this study.

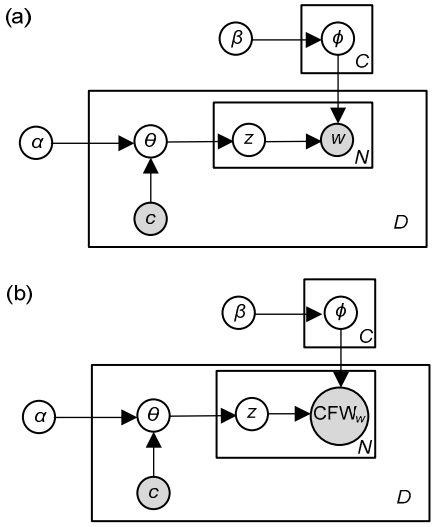

Fig. 1 Models of labeled latent Dirichlet allocation (L-LDA) (a) and weighted L-LDA (WL-LDA) (b)

1. VI

The basic idea of VI (Blei et al., 2003) is to use Jensen's inequality to approach the tightest lower bound on the log likelihood. Toward this goal, it introduces a variational distribution $q(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{z} \mid \widehat{\beta}, \widehat{\alpha}, \widehat{\theta})$:

$$q(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{z} \mid \widehat{\beta},\widehat{\alpha},\widehat{\theta}) = \prod_{c=1}^{C} q(\boldsymbol{\phi}_c \mid \widehat{\beta}_c) \cdot \prod_{d=1}^{D}\left(q(\boldsymbol{\theta}_d \mid \widehat{\alpha}_d) \prod_{n=1}^{N_d} q(z_{d,n} \mid \hat{\theta}_{d,n}) \right), \tag{3}$$

where $\widehat{\beta}$ and $\widehat{\alpha}$ are Dirichlet parameters, and $\hat{\theta}$ is the multinomial distribution parameter. The task of finding the tightest lower bound on the log likelihood can be transformed to maximize the following lower bound:

$$L(\widehat{\beta},\widehat{\alpha},\hat{\theta} \mid \boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{z}) = E_q\left[\log p(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{z} \mid W,\alpha,\boldsymbol{\beta})\right] - E_q\left[\log q(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{z} \mid \widehat{\beta},\widehat{\alpha},\hat{\theta})\right], \tag{4}$$

where $E_q[\cdot]$ is the expectation with respect to the variational distribution. A coordinate ascent algorithm is used to maximize the function L with respect to the variational parameters $\widehat{\beta}$, $\widehat{\alpha}$, and $\hat{\theta}$. For each document d with the label set y_d, two variational parameters $\widehat{\alpha}$ and $\hat{\theta}$ are updated as follows:

$$\begin{cases} \hat{\theta}_{d,n,c} \propto \exp\left(\Psi(\widehat{\alpha}_{d,c}) + \Psi(\widehat{\beta}_{c,w_{d,n}}) - \Psi\sum_{v=1}^{V} \widehat{\beta}_{c,v} \right), \\ \widehat{\alpha}_{d,c} = \alpha + \sum_{n=1}^{N_d} \mathrm{CFW}_{w_{d,n}} \hat{\theta}_{d,n,c}, \quad \text{s.t. } c \in y_d, \end{cases} \tag{5}$$

where $\Psi(\cdot)$ is the digamma function. After the optimal $\widehat{\alpha}$ and $\hat{\theta}$ are obtained for all the documents, the variational parameter $\widehat{\beta}$ is updated as follows:

$$\widehat{\beta}_{c,v} = \beta_v + \sum_{d=1}^{D}\sum_{n=1}^{N_d} \mathrm{CFW}_{w_{d,n}} \hat{\theta}_{d,n,c} f(w_{d,n}), \tag{6}$$

where $f(w_{d,n})$ is equal to 1 or 0 depending on whether $w_{d,n}$ is equal to v or not. Finally, the label-word distribution $\boldsymbol{\phi}$ is calculated for each label c as

$$\phi_{c,v} = \frac{\widehat{\beta}_{c,v}}{\sum_{v=1}^{V} \widehat{\beta}_{c,v}}. \tag{7}$$

The VI algorithm for WL-LDA is outlined in Algorithm 3.

Algorithm 3 VI for WL-LDA

Initialize: parameters
repeat
  for d=1, 2, …, D **do**
    $\forall\, c \in y_d$, initialize $\hat{\alpha}_{d,c} = \alpha + \sum_{n=1}^{N_d} \mathrm{CFW}_{w_{d,n}} / |y_d|$ and $\hat{\theta}_{d,n,c} = 1/|y_d|$
    repeat
      Compute $\hat{\alpha}_d$ and $\hat{\theta}_d$ using Eq. (5)
    until convergence
  end for
  Compute $\hat{\beta}$ using Eq. (6)
until convergence
Compute all C $\boldsymbol{\phi}$ using Eq. (7)

2. Gibbs sampler

The Gibbs sampler is a kind of Markov chain Monte Carlo (MCMC) algorithm. It trains topic models (Griffiths and Steyvers, 2004) by sequentially updating the latent assignment $z_{d,n}$ for each word token. For our WL-LDA model, because the label sets of the training data are observed, the update rule of

$z_{d,n}$ is derived as follows:

$$P(z_{d,n}=c\,|\,W,\boldsymbol{z}_d^{-n},\alpha,\beta)\propto\frac{\mathrm{SN}_{d,c}^{-n}+\alpha}{\mathrm{SN}_d^{-n}+|y_d|\alpha}\cdot\frac{\mathrm{SN}_{c,w_{d,n}}^{-n}+\beta}{\mathrm{SN}_c^{-n}+V\beta}\quad\forall\ c\in y_d,\tag{8}$$

where $\mathrm{SN}_{d,c}$ and SN_d are the number of words with the CF-weight assigned to label c and the total number of words with the CF-weight in document d, respectively, $\mathrm{SN}_{c,w_{d,n}}$ and SN_c are the number of words v with the CF-weight assigned to label c and the total number of words with the CF-weight assigned to label c, respectively, and the superscript '$-n$' is a quantity except for the word token in position n. Given the burn-in samples, we can obtain a point estimate of the label-word distribution $\boldsymbol{\phi}$ for each label c as

$$\phi_{c,v}=\frac{\mathrm{SN}_{c,v}+\beta}{\mathrm{SN}_c+V\beta}.\tag{9}$$

The Gibbs sampler for WL-LDA is outlined in Algorithm 4.

Algorithm 4 Gibbs sampler for WL-LDA

```
for each document d do
  Randomly initialize the assignment z_d according to y_d
  repeat
    for d=1, 2, …, D do
      Update z_d using Eq. (8)
    end for
  until convergence
  Compute all C φ using Eq. (9)
end for
```

3.1.2 Inference for test documents

During the testing, WL-LDA predicts unseen documents by their document-label distributions $\boldsymbol{\theta}$. Because the labels of text documents are unknown, each test document is free to sample all the labels. Suppose that $\boldsymbol{\phi}^*$ is the optimal label-word distribution obtained in the training procedure.

For VI, each test document d is estimated as follows:

$$\begin{cases}\hat{\theta}_{d,n,c}\propto\phi^*_{c,w_{d,n}}\exp[\Psi(\hat{\alpha}_{d,c})],\\ \hat{\alpha}_{d,c}=\alpha+\sum_{n=1}^{N_d}(\mathrm{CFW}_{w_{d,n}}\hat{\theta}_{d,n,c}).\end{cases}\tag{10}$$

The corresponding document-label distribution $\boldsymbol{\theta}_d$ can be obtained by

$$\theta_{d,c}=\frac{\hat{\alpha}_{d,c}}{\sum_{i=1}^{c}\hat{\alpha}_{d,i}}.\tag{11}$$

For the Gibbs sampler, each test document d is estimated as follows:

$$P(z_{d,n}=c\,|\,d,\boldsymbol{z}_d^{-n},\alpha,\boldsymbol{\phi}^*)\propto\phi^*_{c,w_{d,n}}\frac{\mathrm{SN}_{d,c}^{-n}+\alpha}{\mathrm{SN}_d^{-n}+C\alpha}.\tag{12}$$

The corresponding document-label distribution $\boldsymbol{\theta}_d$ can be obtained by

$$\theta_{d,c}=\frac{\mathrm{SN}_{d,c}+\alpha}{\mathrm{SN}_d+C\alpha}.\tag{13}$$

3.2 Weighting the words for dependency-LDA

CF-weight is used for dependency-LDA, and then a weighted dependency-LDA (WD-LDA) is proposed. Again, in WD-LDA, each word token is weighted by its corresponding CF-weight.

3.2.1 Model training

The goal of model training for WD-LDA is to estimate all C label-word distributions $\boldsymbol{\phi}$ and all T topic-label distributions $\boldsymbol{\varphi}$. Given an observed data W, it is achieved by approximating the posterior distribution of the latent variables $p(\boldsymbol{\varphi}, \boldsymbol{\phi}, \boldsymbol{\theta}', \boldsymbol{\theta}, \boldsymbol{z}', \boldsymbol{c}', \boldsymbol{z}|W, \boldsymbol{\beta}, \boldsymbol{\beta}', \boldsymbol{\alpha}, \boldsymbol{\alpha}')$. In this study, the Gibbs sampler is used to learn WD-LDA. Following the original dependency-LDA model, we declare some settings for the training: (1) Restricting the training document to its own label set, the parameter α_d is set to 0; (2) In Eq. (1), M_d is set to the value of $|y_d|$, and $N_{d,i}$ is set to $1/|y_d|$ or 0 depending on whether $i\in y_d$ or not; (3) For each document d, we randomly assign the upper label tokens' $\boldsymbol{c}_d'$ labels included in y_d and fix $\boldsymbol{c}_d'$ during training.

Under the settings suggested above, the asymmetric Dirichlet prior $\boldsymbol{\alpha}_d$ is in fact fixed; thus, the distributions $\boldsymbol{\phi}$ and $\boldsymbol{\varphi}$ are conditionally independent of each other. For the label-word distribution $\boldsymbol{\phi}$, we sequentially update the label assignment of each $z_{d,n}$ word token as follows:

$$P(z_{d,n}=c\,|\,W,\boldsymbol{z}_d^{-n},\boldsymbol{\alpha}_d,\beta)\propto\frac{\mathrm{SN}_{d,c}^{-n}+\alpha_{d,c}}{\mathrm{SN}_d^{-n}+\sum_{i=1}^{C}\alpha_{d,i}}\cdot\frac{\mathrm{SN}_{c,w_{d,n}}^{-n}+\beta}{\mathrm{SN}_c^{-n}+V\beta},\forall\, c\in y_d. \quad (14)$$

For the topic-label distribution $\boldsymbol{\varphi}$, because $\boldsymbol{c}'_d$ is fixed, we can sequentially update each $z'_{d,i}$ of the M_d topic assignments as follows:

$$P(z'_{d,i}=t\,|\,\boldsymbol{c}_d'^{-i},\boldsymbol{z}'_d,\alpha',\beta')\propto(N_{d,t}^{-i}+\alpha')\cdot\frac{N_{t,c'_{d,i}}^{-i}+\beta'}{N_t^{-i}+C\beta'}, \quad (15)$$

where $N_{t,c}$ is the number of times that label c has been assigned to topic t, and N_t is the total number of labels assigned to topic t.

Given the burn-in samples, the label-word distribution $\boldsymbol{\phi}$ is estimated by Eq. (7), and the topic-label distribution $\boldsymbol{\varphi}$ is estimated as follows:

$$\varphi_{t,c}=\frac{N_{t,c}+\beta'}{N_t+C\beta'}. \quad (16)$$

3.2.2 Inference for test documents

During the testing, given the optimal distributions $\boldsymbol{\phi}^*$ and $\boldsymbol{\varphi}^*$ obtained from the training procedure, we want to infer the document-label distributions $\boldsymbol{\theta}$ for unseen documents. Because the labels for test documents are unknown, each test document is free to sample all the labels. The inference for WD-LDA is quite fussy and inefficient. To this end, the fast approximate Gibbs sampler inference method (Rubin et al., 2012) is used in our model. Suppose that M_d=N_d for each test document d, we state a full cycle update process for this fast inference method as follows:

1. Update the label assignment of word tokens to one of the C labels by

$$P(z_{d,n}=c\,|\,d,\boldsymbol{z}_d^{-n},\boldsymbol{\alpha}_d,\boldsymbol{\phi}^*)\propto\phi^*_{c,w_{d,n}}\frac{\mathrm{SN}_{d,c}^{-n}+\alpha_{d,c}}{\mathrm{SN}_d^{-n}+\sum_{i=1}^{C}\alpha_{d,i}}. \quad (17)$$

2. Set the upper label tokens as equal to the current label assignments $\boldsymbol{c}'_d$=$\boldsymbol{z}_d$.

3. Update the topic assignment of the upper label tokens to one of the T topics by

$$P(z'_{d,i}=t\,|\,\boldsymbol{c}_d'^{-i},\boldsymbol{z}'_d,\alpha',\boldsymbol{\varphi}^*)\propto\varphi^*_{t,c'_{d,i}}(N_{d,t}^{-i}+\alpha'). \quad (18)$$

4. Compute the asymmetric document-label Dirichlet prior $\boldsymbol{\alpha}_d$ by

$$\alpha_{d,c}=\eta_d\sum_{t=1}^{T}\varphi^*_{t,c}\cdot\theta'_{d,t}+\alpha_d, \quad (19)$$

where $\theta'_{d,t}$=$(N_{d,t}+\alpha')/(M_d+T\alpha')$.

Finally, each test document d is predicted by estimating the posterior distribution $\boldsymbol{\theta}_d$ over labels:

$$\theta_{d,c}=\frac{\mathrm{SN}_{d,c}+\alpha_{d,c}}{\mathrm{SN}_d+\sum_{i=1}^{C}\alpha_{d,i}}. \quad (20)$$

3.3 Related work

There are some works that weighted words in topic modeling (Debole and Sebastiani, 2004; Madsen et al., 2005; Guan et al., 2009; Petterson et al., 2010; Reisinger et al., 2010; Wilson and Chew, 2010; Shang et al., 2011; Lee et al., 2015; Li et al., 2015a). They used mainly traditional word weighting schemes like term frequency-inverse document frequency (TF-IDF) and pairwise mutual information (PMI). A spherical topic model (Reisinger et al., 2010) uses TF-IDF directly to weight all words and a similar word weighting topic model can be found in Madsen et al. (2005). The model proposed by Wilson and Chew (2010) computes document-level PMI values as word weights. In contrast to these models, the focus of our model is on multi-label classification, and we weight words by the class frequency information, which is a more discriminative knowledge in classification (Guan et al., 2009; Li et al., 2015a). In addition, a recent topic model (Lee et al., 2015) considers word weights by introducing the 'variance' among topic-word distributions in the model generative process, and then uses the topic proportions of documents as features for document classification. The empirical results showed that the model achieved a competitive classification performance. However, it is a little bit time-consuming since it introduces an additional generative step. In our models, only the simpler single-labeled setting is considered.

4 Experiment

4.1 Experimental settings

4.1.1 Dataset

The proposed models were evaluated on seven commonly used multi-label datasets (Machine Learning & Knowledge Discovery Group, 2011), including medical, enron, rcv1subset1, bibtex, a subset of bookmarks, and two subdirectory datasets of Yahoo! (arts and health). The statistics of these datasets are outlined in Table 2, including the numbers of documents and labels, and the cardinality (i.e., the average number of labels per document).

Table 2 Statistics of the datasets used

Dataset	Number of documents	Number of labels	Cardinality
Yahoo! arts	7484	26	1.7
Yahoo! health	9205	32	1.6
Medical	978	45	1.2
Enron	1694	53	3.4
Rcv1subset1	6000	101	2.9
Bibtex	7395	159	2.4
Bookmarks	15 000	208	2.0

4.1.2 Baseline algorithm

Four baseline algorithms were used: two supervised topic-modeling algorithms, L-LDA and dependency-LDA, and two discriminative algorithms, support vector machines (SVMs) and random k-labelsets (RAkLE) (Tsoumakas et al., 2011b). For the two discriminative algorithms, the normalized TF-IDF representation (Salton and Buckley, 1988) was used to encode the documents.

The settings of all these baseline algorithms are given as follows:

1. For L-LDA, we implemented an in-house code using the Gibbs sampler. In our experiments, the symmetric Dirichlet priors were used, where α=50/C and β=0.01.

2. For dependency-LDA, we implemented an in-house code for the fast inference method. We tuned all the parameters according to the suggestions from Rubin et al. (2012). The final parameter settings are shown in Table 3.

3. For SVMs, we used the popular LibSVM tool (Chang and Lin, 2016) and tuned its parameters by a linear search on the set $\{2^i|i=-5, -4, \ldots, 4, 5\}$.

4. For RAkLE, the well-known Mulan tool (Tsoumakas et al., 2011a) was employed. Following the suggestions from Tsoumakas et al. (2011b), the size of the labelsets was set to 3, the number of labelsets was set to 2C, and the C4.5 decision tree was chosen as the base-level algorithm.

4.1.3 Evaluation metric

Two popular evaluation methods for multi-label classification, the Micro-F1 and Macro-F1, were used in our experiments. The F1 metric is the harmonic mean of precision and recall. Given the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), the F1 metric is obtained by

$$\mathrm{F1}=\frac{2\cdot\mathrm{TP}}{2\cdot\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}. \tag{21}$$

Micro-F1 and Macro-F1 are the micro- and macro-averaged versions with respect to the F1 metric, respectively. They are used to measure the binary prediction performance across labels. Define a binary evaluation measure B(TP, FP, TN, FN). Let TP_c, FP_c,

Table 3 Parameter settings for dependency-LDA

Dataset	Model training						Testing		
	η_d	α_d	β	T	β'	α'	η_d	α_d (sum)	α' (sum)
Yahoo! arts	50	0	0.01	20	1	0.01	100	1	1
Yahoo! health	50	0	0.01	20	1	0.01	100	1	1
Medical	50	0	0.01	100	10	0.01	100	1	10
Enron	50	0	0.01	100	10	0.01	100	1	10
Rcv1subset1	50	0	0.01	100	10	0.01	100	1	10
Bibtex	50	0	0.01	200	10	0.01	100	1	10
Bookmarks	50	0	0.01	200	10	0.01	100	1	10

TN_c, and FN_c be the numbers of true positives, false positives, true negatives, and false negatives after binary evaluation in terms of the label c, respectively. The Micro-F1 and Macro-F1 metrics can be computed as follows:

$$\text{Micro-F1} = B\left(\sum_{c=1}^{C}\text{TP}_c, \sum_{c=1}^{C}\text{FP}_c, \sum_{c=1}^{C}\text{TN}_c, \sum_{c=1}^{C}\text{FN}_c\right), \quad (22)$$

$$\text{Macro-F1} = \frac{1}{C}\sum_{c=1}^{C} B\left(\text{TP}_c, \text{FP}_c, \text{TN}_c, \text{FN}_c\right). \quad (23)$$

Both of them suffice for the higher the better.

4.2 Comparison with baseline algorithms

For WL-LDA, we set the two parameters in Eq. (2) as η=2 and γ=0.01. The other parameters were set the same as those of L-LDA. For clarity, WL-LDA estimated by VI and the Gibbs sampler are called WL-LDA$_V$ and WL-LDA$_G$, respectively. For WD-LDA, we set the two parameters in Eq. (2) as η=3 and γ=0.01. The other parameters were set the same as those of dependency-LDA.

For all datasets, we generated repeatedly 10 splits, where each split used 66% of the documents as the training data and the rest as the test data. The

Table 4 Experimental performance of Micro-F1

Dataset	Micro-F1						
	WL-LDA$_V$	WL-LDA$_G$	L-LDA	WD-LDA	Dep-LDA	SVMs	RAkLE
Yahoo! arts	0.451 ±0.0039	0.437 ±0.0019	0.412 ±0.0059	**0.468 ±0.0014**	0.454 ±0.0031	0.435 ±0.0025	0.446 ±0.0121
Yahoo! health	0.603 ±0.0058	0.591 ±0.0096	0.578 ±0.0129	**0.626 ±0.0073**	0.616 ±0.0098	0.617 ±0.0132	0.614 ±0.0095
Medical	0.795 ±0.0108	0.782 ±0.0091	0.769 ±0.0139	**0.805 ±0.0125**	0.792 ±0.0174	0.791 ±0.0102	0.793 ±0.0121
Enron	0.393 ±0.0168	0.401 ±0.0124	0.381 ±0.0228	0.523 ±0.0114	0.514 ±0.0132	0.526 ±0.0183	**0.554 ±0.0091**
Rcv1subset1	0.257 ±0.0034	0.248 ±0.0085	0232 ±0.0101	**0.269 ±0.0028**	0.248 ±0.0084	0.258 ±0.0045	0.239 ±0.0095
Bibtex	0.376 ±0.0065	0.369 ±0.0103	0.361 ±0.0085	**0.411 ±0.0021**	0.375 ±0.0025	0.398 ±0.0046	0.404 ±0.0068
Bookmarks	0.225 ±0.0113	0.218 ±0.0057	0.193 ±0.0124	**0.239 ±0.0065**	0.211 ±0.0095	0.212 ±0.0086	0.216 ±0.0093

Dep-LDA is the abbreviation for dependence-LDA. The best results of each dataset are in boldface

Table 5 Experimental performance of Macro-F1

Dataset	Macro-F1						
	WL-LDA$_V$	WL-LDA$_G$	L-LDA	WD-LDA	Dep-LDA	SVMs	RAkLE
Yahoo! arts	0.302 ±0.0017	0.306 ±0.0033	0.287 ±0.0014	**0.322 ±0.0029**	0.318 ±0.0032	0.301 ±0.0053	0.314 ±0.0025
Yahoo! health	0.271 ±0.0051	0.269 ±0.0067	0.218 ±0.0083	**0.319 ±0.0153**	0.310 ±0.0143	0.288 ±0.0072	0.289 ±0.0059
Medical	0.341 ±0.0242	0.334 ±0.0153	0.329 ±0.0295	0.356 ±0.0192	0.324 ±0.0348	0.358 ±0.0246	**0.369 ±0.0183**
Enron	0.123 ±0.0097	0.127 ±0.0137	0.094 ±0.0101	**0.149 ±0.0057**	0.113 ±0.0062	0.142 ±0.0089	0.147 ±0.0064
Rcv1subset1	0.136 ±0.0057	0.132 ±0.0072	0.131 ±0.0094	**0.142 ±0.0068**	0.133 ±0.0102	0.139 ±0.0088	0.137 ±0.0053
Bibtex	0.272 ±0.0062	0.283 ±0.0048	0.221 ±0.0083	**0.301 ±0.0037**	0.293 ±0.0049	0.282 ±0.0068	0.268 ±0.0054
Bookmarks	0.114 ±0.0086	0.115 ±0.0052	0.098 ±0.0136	**0.132 ±0.0084**	0.112 ±0.0102	0.082 ±0.0077	0.088 ±0.0126

Dep-LDA is the abbreviation for dependence-LDA. The best results of each dataset are in boldface

mean and standard deviation values of all the algorithms were then reported.

Experimental results are shown in Tables 4 and 5. Overall, we observe that WD-LDA outperforms the other algorithms in most cases, and that our weighted models perform better than the non-weighted versions. Further discussions are given as follows:

1. WL-LDA vs. L-LDA

First, we see that the performance of WL-LDA is higher than that of L-LDA on all the datasets. For the datasets with fewer labels, i.e., Yahoo! arts, Yahoo! health, medical, and enron, WL-LDA achieves about 0.01–0.04 Micro-F1 improvements and about 0.01–0.05 Macro-F1 improvements over L-LDA. For other datasets with more labels, the performance gaps between WL-LDA and L-LDA are relatively small, e.g., about 0.01 on Micro-F1 of the bibtex dataset and about 0.005 on Macro-F1 of the rcv1subset1 dataset. The only exception is the gap on Macro-F1 of the bibtex dataset (0.06). This indicates that our CF-weight algorithm can improve the classification accuracy for rare labels. Second, for the two WL-LDA versions, WL-LDA$_V$ performs better (6/7 datasets on Micro-F1 and 3/7 datasets on Macro-F1) than WL-LDA$_G$, where the largest gap is 0.014.

2. WD-LDA vs. Dependency-LDA

Obviously, WD-LDA outperforms the non-weighted dependency-LDA on both Micro-F1 and Macro-F1 across all the datasets. The improvements are stable on different types of datasets, i.e., about 0.004–0.036. For the datasets with more labels, the improvements of WD-LDA are larger, e.g., about 0.036 on Micro-F1 across the bibtex dataset and about 0.02 on Macro-F1 across the bookmarks dataset. These results indicate that using the CF-weight improves successfully the multi-label classification performance of the dependency-LDA.

3. CF-weight based models vs. SVMs and RAkLE

First, the simpler WL-LDA model can achieve competitive performance with SVMs (6/14 datasets) and RAkLE (6/14 datasets). The two discriminative algorithms use the TF-IDF method to represent documents. In this sense, they have considered the document frequency information, i.e., the number of documents where the word has occurred. Both class frequency and document frequency focus on the universality of words. This is the main reason that SVMs and RAkLE outperform L-LDA, but they are almost competitive with WL-LDA. Second, WD-LDA performs better than SVMs and RAkLE. Compared to SVMs, WD-LDA achieves about 0.003–0.05 improvements on most of the settings (12/14 datasets), and its performance is a bit lower on the Micro-F1 (0.003) across the enron dataset and on Macro-F1 (0.002) across the medical dataset. Compared to RAkLE, WD-LDA also achieves improvements on most of the settings (12/14 datasets), e.g., about 0.030 on Micro-F1 across the rcv1subset1 dataset and about 0.030 on Macro-F1 across the health dataset. We argue that the lead performance of WD-LDA is due to the fact that WD-LDA considers the class frequency of the word, label frequency, and label dependency at the same time, but other algorithms consider only some of these attributes.

4. Comparisons with discriminate algorithms with feature selection

Since the CF-weight scheme can be deemed a word (i.e., feature) selection in supervised topic models, we further compared our algorithms with discriminate algorithms with feature selection. Two feature selection methods for classification, information gain and Chi-square, were used. We reported only the better Micro-F1 scores between the two feature selection methods since Macro-F1 scores are very close to Micro-F1 ones, and compared them with those of WD-LDA. The results are shown in Table 6. We can see that WD-LDA outperforms the two discriminative algorithms with feature selection in most of the settings. This indicates that the CF-weight has a greater influence on positive classification.

Table 6 Comparisons between WD-LDA and discriminative algorithms (SVMs and RAkLE) with feature selection on Micro-F1

Dataset	Micro-F1		
	WD-LDA	SVMs	RAkLE
Yahoo! arts	**0.468**	0.441	0.441
	±0.0014	±0.0022	±0.0095
Yahoo! health	**0.626**	0.616	0.615
	±0.0073	±0.0106	±0.0108
Medical	**0.805**	0.799	0.796
	±0.0125	±0.0082	±0.0151
Enron	0.523	0.537	**0.551**
	±0.0114	±0.0205	**±0.0084**
Rcv1subset1	**0.269**	0.261	0.247
	±0.0028	±0.0037	±0.0107
Bibtex	**0.411**	0.402	0.407
	±0.0021	±0.0063	±0.0054
Bookmarks	**0.239**	0.217	0.219
	±0.0065	±0.0059	±0.0072

The best results of each dataset are in boldface

4.3 Study on the parameters

The scale coefficient η of CF-weight was investigated for WL-LDA and WD-LDA. In the experiments, η was evaluated with different values from the set {1, 2, …, 8}. Since early experimental results indicated that the performances of Micro-F1 and Macro-F1 have very similar trends, we reported only the Micro-F1 scores. Four datasets with different statistics were used: the Yahoo! arts dataset with fewer labels and smaller cardinality, the enron dataset with fewer labels and larger cardinality, the rcv1subset1 dataset with more labels and larger cardinality, and the bookmarks dataset with more labels and lower cardinality.

Experimental results are shown in Table 7. Overall, we can observe that the Micro-F1 scores are higher when η is relatively small, and they go down as η increases. For example, the Micro-F1 performance gap between η=8 and η=2 is about 0.05–0.1. The possible reason is as follows: In our algorithm, the parameter η is used to control the importance of the CF-weight; i.e., the CF-weight becomes more important as η increases. We may lose other useful knowledge like word co-occurrences if we overemphasize the CF-weight. Additionally, we observe that the best scores are around η=2 for WL-LDA and around η=3 for WD-LDA. The optimal setting of η in WD-LDA is larger than that in WL-LDA. This is because during model training, the asymmetric document-label Dirichlet prior for WD-LDA is often very large, dominating the document-label inference. It needs larger η values to enlarge the CF-weights. In practice, we suggest η=2 or 3 as the default settings in our weighted supervised topic models.

5 Conclusions

In this paper, we have investigated the problem that the existing supervised topic models neglect the class frequencies of words, which is significant for classification. To this end, we have suggested a weighted word method based on class frequency knowledge, namely CF-weight, which provides higher weights for the words with small class frequency. We have considered the CF-weight to be the soft count of words, and then used it to improve two of the state-of-the-art supervised topic models L-LDA and dependency-LDA. Extensive experiments have been conducted to evaluate our algorithms. The empirical results validated that the proposed CF-weight is beneficial to both L-LDA and dependency-LDA, and it can achieve competitive multi-label classification performance.

References

Blei DM, McAuliffe JD, 2007. Supervised topic models. 20th Int Conf on Neural Information Processing Systems, p.121-128.

Blei DM, Ng AY, Jordan MI, 2003. Latent Dirichlet allocation. *J Mach Learn Res*, 3:993-1022.

Chang CC, Lin CJ, 2016. LIBSVM—a Library for Support Vector Machines. https://www.csie.ntu.edu.tw/~cjlin/libsvm/ [Accessed on May 22, 2018].

Table 7 Micro-F1 performance with different values of the scale coefficient η

Model	Dataset	Micro-F1							
		η=1	η=2	η=3	η=4	η=5	η=6	η=7	η=8
$WL\text{-}LDA_V$	Yahoo! arts	**0.452**	0.451	0.449	0.437	0.422	0.414	0.402	0.375
	Enron	0.384	**0.393**	0.387	0.369	0.361	0.332	0.308	0.303
	Rcv1subset1	0.249	**0.257**	0.253	0.246	0.205	0.194	0.157	0.176
	Bookmarks	0.221	**0.225**	**0.225**	0.214	0.219	0.184	0.173	0.159
$WL\text{-}LDA_G$	Yahoo! arts	0.433	**0.437**	0.436	0.421	0.423	0.402	0.397	0.385
	Enron	0.394	**0.401**	0.399	0.379	0.352	0.345	0.321	0.317
	Rcv1subset1	0.239	0.248	**0.251**	0.242	0.225	0.205	0.167	0.153
	Bookmarks	0.217	0.219	**0.221**	0.217	0.202	0.191	0.169	0.148
WD-LDA	Yahoo! arts	0.459	0.462	**0.468**	0.451	0.439	0.407	0.376	0.394
	Enron	**0.525**	0.517	0.523	0.509	0.476	0.452	0.467	0.426
	Rcv1subset1	0.263	**0.269**	**0.269**	0.264	0.209	0.204	0.217	0.175
	Bookmarks	0.211	0.225	**0.239**	0.217	0.203	0.167	0.158	0.136

The best results of each dataset are in boldface

Debole F, Sebastiani F, 2004. Supervised term weighting for automated text categorization. In: Sirmakessis S (Ed.), Text Mining and Its Applications. Springer, Berlin, p.81-97. https://doi.org/10.1007/978-3-540-45219-5_7

Ghahramani Z, 2001. An introduction to hidden Markov models and Bayesian networks. *Int J Patt Recogn Artif Intell*, 15(1):9-42. https://doi.org/10.1142/S0218001401000836

Griffiths TL, Steyvers M, 2004. Finding scientific topics. *Proc Nat Acad Sci USA*, 101(Suppl 1):5228-5235. https://doi.org/10.1073/pnas.0307752101

Guan H, Zhou JY, Guo MY, 2009. A class-feature-centroid classifier for text categorization. 18th Int Conf on World Wide Web, p.201-210. https://doi.org/10.1145/1526709.1526737

Kim D, Kim S, Oh A, 2012. Dirichlet process with mixed random measures: a nonparametric topic model for labeled data. 29th Int Conf on Machine Learning, p.675-682.

Lacoste-Julien S, Sha F, Jordan MI, 2008. DiscLDA: discriminative learning for dimensionality reduction and classification. 21st Int Conf on Neural Information Processing Systems, p.897-904.

Lee S, Kim J, Myaeng SH, 2015. An extension of topic models for text classification: a term weighting approach. Int Conf on Big Data and Smart Computing, p.217-224. https://doi.org/10.1109/35021BIGCOMP.2015.7072834

Li XM, Ouyang JH, Zhou XT, 2015a. Centroid prior topic model for multi-label classification. *Patt Recogn Lett*, 62:8-13. https://doi.org/10.1016/j.patrec.2015.04.012

Li XM, Ouyang JH, Zhou XT, 2015b. Supervised topic models for multi-label classification. *Neurocomputing*, 149:811-819. https://doi.org/10.1016/j.neucom.2014.07.053

Machine Learning & Knowledge Discovery Group, 2011. Learning from Multi-label Data. http://mlkd.csd.auth.gr/multilabel.html [Accessed on May 12, 2018].

Madsen RE, Kauchak D, Elkan C, 2005. Modeling word burstiness using the Dirichlet distribution. 22nd Int Conf on Machine Learning, p.545-552. https://doi.org/10.1145/1102351.1102420

Petterson J, Smola A, Caetano T, et al., 2010. Word features for latent Dirichlet allocation. 23rd Int Conf on Neural Information Processing Systems, p.1921-1929.

Ramage D, Hall D, Nallapati R, et al., 2009. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. Conf on Empirical Methods in Natural Language Processing, p.248-256. https://doi.org/10.3115/1699510.1699543

Ramage D, Manning CD, Dumais S, 2011. Partially labeled topic models for interpretable text mining. 17th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining, p.457-465. https://doi.org/10.1145/2020408.2020481

Reisinger J, Waters A, Silverthorn B, et al., 2010. Spherical topic models. Proc 27th Int Conf on Machine Learning, p.1-8.

Rubin TN, Chambers A, Smyth P, et al., 2012. Statistical topic models for multi-label document classification. *Mach Learn*, 88(1-2):157-208. https://doi.org/10.1007/s10994-011-5272-5

Salton G, Buckley C, 1988. Term-weighting approaches in automatic text retrieval. *Inform Process Manag*, 24(5): 513-523. https://doi.org/10.1016/0306-4573(88)90021-0

Shang LF, Chan KP, Pan GD, 2011. DTTM: a discriminative temporal topic model for facial expression recognition. 7th Int Conf on Advances in Visual Computing, p.596-606. https://doi.org/10.1007/978-3-642-24028-7_55

Tsoumakas G, Spyromitros-Xioufis E, Vilcek J, et al., 2011a. Mulan: a Java library for multi-label learning. *J Mach Learn Res*, 12(7):2411-2414.

Tsoumakas G, Katakis I, Vlahavas I, 2011b. Random *k*-labelsets for multilabel classification. *IEEE Trans Knowl Data Eng*, 23(7):1079-1089. https://doi.org/10.1109/TKDE.2010.164

Wilson AT, Chew PA, 2010. Term weighting schemes for latent Dirichlet allocation. Human Language Technologies: Annual Conf of the North American Chapter of the Association for Computational Linguistics, p.465-473.

Zhu J, Ahmed A, Xing EP, 2012. MedLDA: maximum margin supervised topic models. 26th Annual Int Conf on Machine Learning, p.1257-1264. https://doi.org/10.1145/1553374.1553535