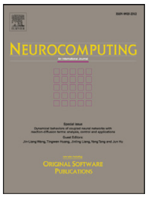
Neurocomputing 312 (2018) 324–335

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Supervised latent Dirichlet allocation with a mixture of sparse softmax

Xiaoxu Li[a,b,*], Zhanyu Ma[a], Pai Peng[c], Xiaowei Guo[c], Feiyue Huang[c], Xiaojie Wang[d], Jun Guo[a]

[a] *School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[b] *School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China*
[c] *YoutuLab, Tecent Technology (Shanghai) Co., Ltd, Shanghai 200233 China*
[d] *School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China*

ARTICLE INFO

Article history:
Received 29 November 2016
Revised 6 May 2018
Accepted 18 May 2018
Available online 26 May 2018

Communicated by zhi yong Liu

Keywords:
Supervised topic model
Ensemble classification
Mixture of softmax model
Latent Dirichlet allocation

ABSTRACT

Real data often show that from appearance within-class similarity is relatively low and between-class similarity is relatively high, which could increase the difficulty of classification. To classify this kind of data effectively, we learn multiple classification criteria simultaneously, and make different classification criterion be applied to classify different data for the purpose of relieving difficulty of fitting this kind of data and class label only by using a single classifier. Considering that topic model can learn high-level semantic features of the original data, and that mixture of softmax model is an efficient and effective probabilistic ensemble classification method, we embed a mixture of softmax model into latent Dirichlet allocation model, and propose a supervised topic model, *supervised latent Dirichlet allocation with a mixture of softmax*, and its improved version, *supervised latent Dirichlet allocation with a mixture of sparse softmax*. Next, we give their parameter estimation algorithms based on variational Expectation Maximization (EM) method. Moreover, we give an approximation method to classify unseen data, and analyze the convergence of the parameter estimation algorithm. Finally, we demonstrate the effectiveness of the proposed models by comparing them with some recently proposed approaches on two real image datasets and one text dataset. The experimental results demonstrate the good performance of the proposed models.

1. Introduction

Classification on real data has been a very important and challenging task in both computer vision and data mining. An important challenge is that some real data often show from appearance within-class similarity is relatively low, while between-class similarity is relatively high. One example of image data is shown in Fig. 1. From appearance images in “croquet” and “bocce” class are very similar, while images fr om the same “polo” class are very different. From appearance images in “inside city” and “tall building” class are very similar, while images from the same “high way” class are very different. For another example from text data, a document introducing a film about artificial intelligence belongs to “entertainment” class, and is quite similar to another document introducing the theory of artificial intelligence but belonging to “science and technology” class. While two documents belonging to the “science and technology” class, which are about biology and astronomy respectively, have quite different text words. This paper focuses on classifying this kind of data or data whose subset belongs to this kind of data.

To classify this kind of data accurately, we try to learn multiple classification criteria simultaneously, and make different classification criterion be applied to classify different data so as to relieve difficulty of fitting this kind of data and class label only by using a single classifier. Generally, one single classification criterion is difficult to fit the relationship between a complex image and its class label, while ensemble methods constructing a combined classification criterion could obtain better predictive performance in theory. There are some successful ensemble methods [1] trying to learn combined classification criterion from low-level features, such as boosting and some conditional mixture models [2].

In addition, topic model originated from text processing [3,4] can learn high-level semantic features of the original data by dimensionality reduction. Until now, much works based on topic model for classification has been done [5–7]. And these works can be fitted into one of two main categories: the double-phase and the single-phase. The double-phase methods tries to first learn high-level features based on unsupervised topic model, and the learned semantic features are then fed into classifiers. The single-phase methods attempt to jointly learn high-level features and

[*] Corresponding author at: School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China.
E-mail addresses: lixiaoxu@lut.cn (X. Li), mazhanyu@bupt.edu.cn (Z. Ma).

https://doi.org/10.1016/j.neucom.2018.05.077
0925-2312/© 2018 Published by Elsevier B.V.

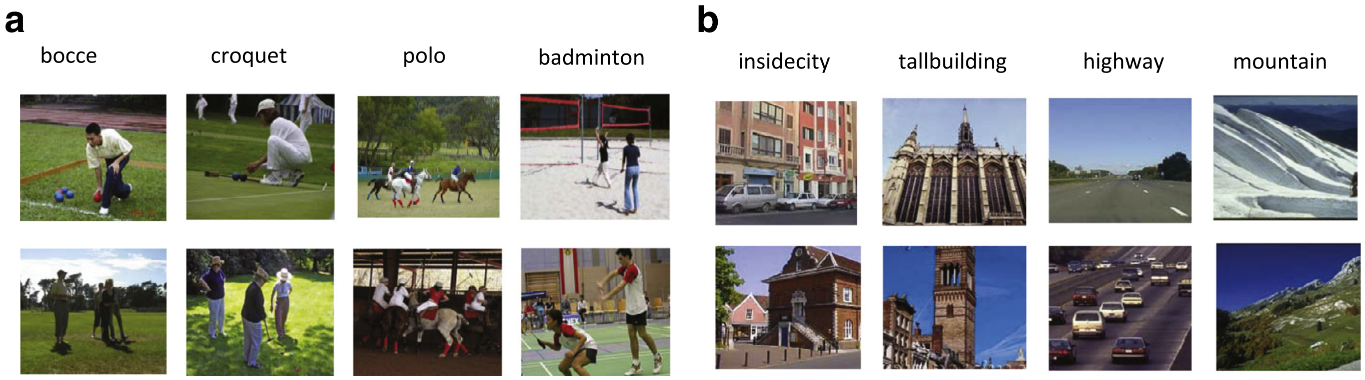


Fig. 1. Example images with the class label from (a). UIUC-Sport dataset [7] (b). LabelMe dataset [11].

classifier, and construct a supervised topic model. A common feature of the two kinds of methods is that all of them construct or learn a single classification criterion (classifier) for all training data. Unlike these works, this paper tries to construct a supervised topic model which has the merits of the ensemble classification.

The paper is built on topic models [8,9] and mixture of softmax model (SMM) [2], in which SMM is a probabilistic ensemble classification method, also constructs a combined classification criterion. We embed a mixture of softmax into LDA model under the framework of supervised LDA, and then propose a supervised topic model, *supervised latent Dirichlet allocation with a mixture of softmax*, and its improved version, *supervised latent Dirichlet allocation with a mixture of sparse softmax*. Moreover, we give a parameter estimation algorithm based on variational EM method [10] and analyze the convergence of the parameter estimation algorithm. In addition, embedding a mixture of softmax into LDA model significantly increases the difficulty in solving the parameters, which is not simply “plug and play”, so that we use some approximation tricks in parameter optimization, as shown in Section 3.

1.1. Related work

Latent aspect models have recently gained much popularity for discovering the semantic aspects (topics) from low-level features of single modal data [3,12] and multi-modal data [6,13,14]. They can be divided into latent aspect models based on directed Bayesian Network and the ones based on undirected Markov network. Undirected Markov network methods, including the exponential family of Harmoniums and its special cases of restricted RBM, enjoys the property of fast inference. In this branch, Chen proposed the infinite exponential family Harmonium (iEFH) [15], a bipartite undirected latent variable model that automatically determines the number of latent units. Zhang et al. [16] developed a new triple wing harmonium Model projecting these multiple textual features into low-dimensional latent topics with different probability distribution assumptions. Compared with undirected Markov network, directed Bayesian Network model the conditional dependencies of variables more directly. In the branch, latent Dirichlet allocation (LDA) model is a classic work. Much extensions and its variants have been developed, including not only general models, such as classification [17,18] and regression [8], but also some variants for some specific applications, such as rating prediction [19] and scene understanding [7,20].

A seminal work for classification based on LDA is the work which was proposed by Fei-Fei and Perona [5]. And it assumed the documents in each category had its own LDA generative process, and each category had its own Dirichlet prior which is optimized and used for distinguishing different classes. Moreover, the Dirichlet priors for categories could be seen as a single classification criterion to a certain extent.

Another similar work, the discriminative LDA (DiscLDA) [21], also built on LDA model, and was trained by maximizing the conditional likelihood of the responses given the documents. The model assumed that the documents in the same category should be near in the obtained topic simplex, and that every category had its own linear transformation, which could transform a point simplex to a new mixture proportion of topics, as well as its own topic distribution.

Supervised latent Dirichlet allocation (sLDA) [8] , another classical topic model, is a supervised extension to LDA model [9]. The model was originally developed for predicting continuous response values via a linear regression, and was trained by maximizing the joint likelihood of data and response variables. Based on framework of sLDA, the literature [17] expanded sLDA to classification problems, and obtained Multi-class sLDA model by embedding softmax regression model into the LDA model. However, Multi-class sLDA model still constructed a single classification criterion.

Maximum entropy discrimination latent Dirichlet allocation model (MedLDA) [22], introduced max-margin idea to LDA model, and proposed a max-margin discriminative variant of supervised topic models for both regression and classification by combining SVM and LDA. The model integrated max-margin learning with hierarchical Bayesian topic models by optimizing a single objective function with a set of expected margin constraints. In addition, considering that some useful image features are not easily represented by “bags of words”, the literature [23] proposed a variant of MedLDA, a max-margin latent Dirichlet allocation (MMLDA) for image classification. The two models based on max-margin idea all assumed a single classification criterion.

Supervised Document Neural Autoregressive Distribution Estimator, SupDocNADE model [24], was different from the models mentioned above, and was a neural topic model based on neural network. The model did not use convolution [25,26], and did not model or define a specific form of topic, but modeled or learnt the map from the low feature to topic feature. The model also assumed a single classification criterion and achieved a competitive performance.

All these existing supervised topic models were trained either by optimizing a likelihood-based objective or by optimizing margin-based objectives. Anyway, they all assumed and learned a single classification criterion. In this paper, we proposed two topic models for classification, which assume there are multiple classification criteria over semantic features. We will introduce them in detail in the following sections, and the remaining sections are organized as follows.

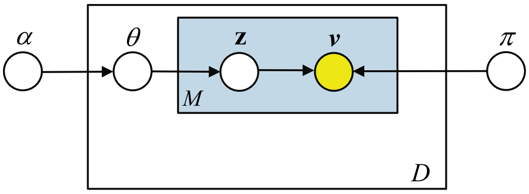


Fig. 2. The graphic model representation of latent Dirichlet allocation model [9].

We introduce the preliminaries of our work in Section 2 firstly. And then, we propose a supervised ensemble topic model and its improved version, and explain how to perform parameter estimation as well as predicting procedure in Section 3. Moreover, we discuss the classification performance of the two proposed models on three datasets in Section 4, and give convergence analysis of parameter estimation procedure in Section 5. Finally, we present our conclusions and future work in Section 6.

Here, it needs to point out that part of the materials of this paper build on our early work [27] which is presented in conference proceedings.

2. Preliminaries

The paper is built on topic models, i.e., LDA, sLDA and mixture of softmax model (SMM) [2]. And these work will be reviewed briefly in this section.

2.1. Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) [9] is a hierarchical Bayesian model that tries to map a text document into a latent low dimensional space spanned by a set of automatically learned topical bases. The model assumes that a document consists K topics, and the generative probability distribution for all the documents is:

$$p(v, z, \theta|\alpha, \pi) = p(\theta|\alpha)\prod_{m=1}^{M} p(z_m|\theta)p(v_m|z_m, \pi) \tag{1}$$

where $\theta \sim \text{Dirichlet}(\alpha)$, which is the topic proportion of a K-dimensional vector and $\sum_{k=1}^{K}\theta_k = 1$. z_m is a K-dimensional indicator vector (only one element is 1, and all others are 0) referring to topic label of the mth word; π is a matrix consisting of K rows for those K topics, with each row representing a multinomial distribution over words in a given vocabulary. The graphical model representation of LDA model is depicted in Fig. 2. The model is often used to find latent high-level features, i.e., topics of a document.

2.2. Supervised latent Dirichlet allocation

As stated above, LDA is an unsupervised topic model, which does not use side information for learning topics and inferring topic vectors θ. For this reason, supervised latent Dirichlet allocation (sLDA) is proposed for modeling data when side information is available. For a document or data, sLDA assumes the same generative process with LDA. The graphical model representation of sLDA model is depicted in Fig. 3. The probability distribution of data under sLDA is shown as follows:

$$\begin{aligned} p(v, y, z, \theta|\alpha, \pi, \sigma^2, \eta) = {} & p(\theta|\alpha)\prod_{m=1}^{M} p(z_m|\theta)p(v_m|z_m, \pi) \\ & \cdot \; p(y|\bar{\boldsymbol{z}}, \sigma^2, \eta) \end{aligned} \tag{2}$$

where $p(\theta|\alpha)\prod_{m=1}^{M} p(z_m|\theta)p(x_m|z_m, \pi)$ represents the same generative process as that of LDA. The supervised information y is subject to normal distribution $N(\eta^\top\bar{\boldsymbol{z}}, \sigma^2)$.

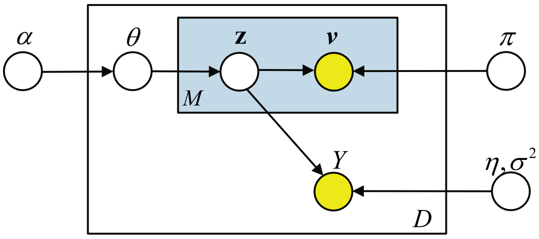


Fig. 3. The graphic model representation of supervised latent Dirichlet allocation model [8].

2.3. Mixture of softmax model

Softmax regression, also known as multinomial logistic regression, is another basis of the paper. The model is a kind of regression model, which generalizes logistic regression by allowing more than two discrete outcomes. Mixture of softmax model [2] is a weighted sum of multiple softmax regression. And it can be seen as a probabilistic ensemble classification model. In particular, the distribution of this model is shown as follows:

$$p(c|x, \eta) = \sum_{h=1}^{H} b_h\left(\frac{\exp(\eta_{hc}^\top x)}{\sum_{l=1}^{C}\exp(\eta_{hl}^\top x)}\right). \tag{3}$$

where x and c represent data and its predicting class label, respectively. H and C are the number of classifiers and classes, respectively. $\eta_{hl}, h \in \{1, 2, \ldots, H\}$ and $l \in \{1, 2, \ldots, C\}$ represents the parameter of the lth class in the hth softmax classifier, and b_h represents the weight of hth softmax classifier. Two well-known improvements to SMM are mixture of experts [2] and hierarchical mixture of experts [2]. In mixture of experts model, the weights of different experts vary over the observed variable space, while hierarchical mixture of experts approach is a mixture of mixtures of experts.

3. Supervised latent Dirichlet allocation with a mixture of sparse softmax

In this section, we will introduce our models which are based on the "bag-of-words" representation as in LDA and sLDA. By this representation, a data is simply reduced to a vector of word counts without considering word order. Given the datasets $\boldsymbol{D} = \{(\boldsymbol{v}_d, c_d)|d \in \{1, 2, \ldots, D\}\}$, our models is to model the relationship between a data point $\boldsymbol{v}$ and its label c.

3.1. Supervised latent Dirichlet allocation with a mixture of softmax

Embedding SMM with L2 regularization into LDA model results in our model, *supervised latent Dirichlet allocation with a mixture of softmax*, or MS-sLDA for short. The model is conditional on two hyper-parameters, H for the number of softmax and K for the number of latent topic for data. For a pair of data-label ($\boldsymbol{v}$, c), in which the label c of a data point $\boldsymbol{v}$ is a unit-basis vector of size C, the generative process of the model is shown as follows:

1. For each data point $\boldsymbol{v} = \{v_m|m \in \{1, 2, \ldots, M\}$:
   (a) Draw a topic proportion $\theta \sim \text{Dirichlet}(\alpha)$.
   (b) For each word v_m, $m \in \{1, 2, \ldots, M\}$:
      i. Draw a topic assignment $z_m = k|\theta \sim \text{Multi}(\theta)$.
      ii. Draw a word $v_m = k|z_m \sim \text{Multi}(\pi_{z_m})$
2. For the class label c:
   (a) Draw a "softmax" assignment $s|\theta^c \sim \text{Multi}(\theta^c)$, where s is a H-dimensional one-hot variable, and its dimension equal the dimension of θ^c and the number of classifiers.

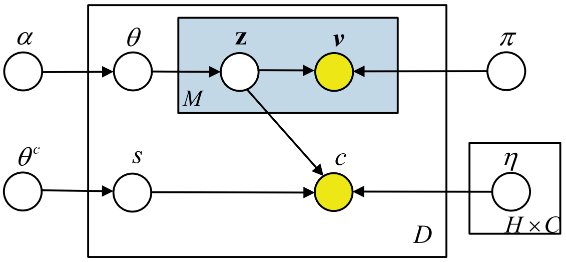


Fig. 4. The graphic model representation of supervised latent Dirichlet allocation with a mixture of softmax.

(b) Draw the class label $c|\boldsymbol{z}$, $s \sim \mathrm{softmax}(\bar{\boldsymbol{z}}, s, \eta)$ where $\bar{\boldsymbol{z}} = \frac{1}{M}\sum_{m=1}^{M}\boldsymbol{z}_m$ is the empirical topic frequencies. The distribution of the class labels can be formulated as:

$$p(c|\bar{\boldsymbol{z}}, s, \eta) = \prod_{h=1}^{H}\left(\frac{\exp(\eta_{hc}^{\top}\bar{\boldsymbol{z}})}{\sum_{l=1}^{C}\exp(\eta_{hl}^{\top}\bar{\boldsymbol{z}})}\right)^{s_h}.$$

The graphical model representation of MS-sLDA model is depicted in Fig. 4. MS-sLDA model specifies a joint distribution over latent variables and observed variables. Denoting model parameters as $\Omega = \{\alpha, \pi, \theta^c, \boldsymbol{\eta}\}$, latent variables as $H = \{\theta, \boldsymbol{z}, s\}$ and observed variables as $E = \{\boldsymbol{v}, c\}$, then the joint distribution is

$$\begin{aligned} p(E, H|\Omega) = {} & p(\theta|\alpha)\prod_{m=1}^{M}p(z_m|\theta)p(v_m|z_m, \pi) \\ & \cdot\ p(s|\theta^c)p(c|\bar{\boldsymbol{z}}, s, \eta) \end{aligned} \tag{4}$$

Note that Step 1, the generative process of a data point $\boldsymbol{v} = \{v_1, v_2, \ldots, v_M\}$, is the same as that of LDA model. Its goal is to obtain the empirical topic frequencies of a data point, which are then used as the input to generate class label in the following step. This step reduces data dimension, so as to reduce the dimension of parameters in softmax classifiers in step 2.

Step 2 modeling the class labels adopts the setup similar to those of sLDA [8] and Multi-class sLDA [17]. Multi-class sLDA considers the class label response variables instead of the continuous response variables in sLDA. And the class label is drawn from a single softmax regression model. That is, the multi-class sLDA model tries to find a single classification criterion (softmax) to identify all training data. While, MS-sLDA model introduces a mixture of softmax model, which can give a higher likelihood comparing to a single softmax, to replace the supervised parts of sLDA. In particular, given the latent topic frequencies $\bar{\boldsymbol{z}}$, the generation of class label needs to draw a latent "criterion" (softmax) assignment $s = h$ first, and then select a label according to the hth "criterion" $\boldsymbol{\eta}_h = (\eta_{h1}, \eta_{h2}, \ldots, \eta_{hC})$.

MS-sLDA model provides an important improvement to sLDA. It can be seen as an ensemble version of Multi-class sLDA model, namely a generalization of the model. When the number of softmax classifiers is set to 1, MS-sLDA model will collapse to the Multi-class sLDA model.

3.2. Parameter estimation

We carry out approximate maximum-likelihood estimation for MS-sLDA model using a variational expectation-maximization (EM) procedure [28,29], which is a normal optimization method taken by many latent topic models, e.g. LDA [9] and sLDA [8].

Variational E-step

Since $p(\theta, \boldsymbol{z}, s, |\boldsymbol{v}, c)$, the posterior distribution of the latent variables conditioned on a pair ($\boldsymbol{v}$, c), is computationally intractable, we adopt variational approximate [10,30,31] to obtain approximate posterior. Given a set of model parameters $\Omega = \{\alpha, \pi, \theta^c, \eta\}$, we find the lower bound of the log likelihood for each data pair ($\boldsymbol{v}$, c) in the training datatset $\boldsymbol{D}$:

$$\begin{aligned} \log p(\boldsymbol{v}, c|\Omega) \geq L(\Lambda; \Omega) = {} & E_q[\log p(\boldsymbol{v}, c, \theta, z, s|\Omega)] \\ & - E_q[\log q(\theta, z, s|\Lambda)], \end{aligned} \tag{5}$$

where $\Lambda = \{\gamma, \boldsymbol{\phi}, \lambda\}$, γ is a K-dimensional Dirichlet parameter, ϕ_m is a K-dimensional multinomial parameter, λ is a H-dimensional multinomial parameter and q is variational distribution over the latent variables for the data ($\boldsymbol{v}$, c). The variational distribution is defined as a full factorized distribution:

$$q(H|\Lambda) = q(\theta|\gamma)\prod_{m=1}^{M}q(z_m|\phi_m)q(s|\lambda), \tag{6}$$

In variational E-step, we maximize the lower bound w.r.t. the variational parameters $\Lambda = \{\gamma, \phi, \lambda\}$, which is equivalent to minimizing the KL-divergence between this factorized distribution and the true posterior.

Opitimization with respect to γ. The procedure is the same as in [9]

$$\gamma_i = \alpha_i + \sum_{m=1}^{M}\phi_{mi}. \tag{7}$$

Opitimization with respect to ϕ_m. The terms including ϕ_m in L are:

$$\begin{aligned} L_{[\phi_m]} = {} & \sum_{i=1}^{K}\phi_{mi}\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k}(\gamma_j)\right) + \sum_{j=1}^{V_s}v_m^j\log\pi_{ij}\right) \\ & + \sum_{h=1}^{H}\lambda_h\left[\eta_{hc}^{\top}\bar{\phi} - \log\left(\sum_{l=1}^{C}E_q\left[\exp(\eta_{hl}^{\top}\bar{\boldsymbol{z}})\right]\right)\right] \\ & - \sum_{i=1}^{k}\phi_{mi}\log\phi_{mi}, \end{aligned}$$

where $\bar{\phi} = \sum_{m=1}^{M}\phi_m/M$. Maximizing the above formulation under the constraint $\sum_{i=1}^{K}\phi_{mi} = 1$ leads to

$$\phi_{mi} \propto \pi_{iv_m}\exp\left[\Psi(\gamma_i) + \sum_{h=1}^{H}\lambda_h\left(\frac{1}{N}\eta_{hci} - (b_h^T\phi_n^{old})^{-1}b_{hi}\right)\right], \tag{8}$$

where $b_{hi} = \sum_{l=1}^{C}\exp(\frac{\eta_{hli}}{M})\prod_{f\neq m}^{M}\left(\sum_{j=1}^{K}\phi_{fi}\exp(\frac{\eta_{hli}}{M})\right)$ and ϕ_n^{old} is the previous value. The detail on computing ϕ_m are similar to [17].

Opitimization with respect to λ. The terms including λ_h in L with approximate Lagrange multipliers are:

$$\begin{aligned} L_{[\lambda_h]} = {} & \sum_{h=1}^{H}\lambda_h\left[\eta_{hc}^{\top}\bar{\phi} + \log(\theta_h^c) - \log\lambda_h\right. \\ & \left. - \log\left(\sum_{l=1}^{C}\prod_{m=1}^{M}\sum_{j=1}^{k}\phi_{mi}\exp\left(\frac{1}{M}\eta_{hlj}\right)\right)\right]. \end{aligned}$$

Maximizing the above formulation under the constraint $\sum_{h=1}^{H}\lambda_h = 1$ leads to

$$\begin{aligned} \lambda_h \propto \exp & \left[\eta_{hc}^{\top}\bar{\phi} - \log\left(\sum_{l=1}^{C}\prod_{m=1}^{M}\sum_{j=1}^{k}\phi_{mi}\exp\left(\frac{1}{M}\eta_{hlj}\right)\right)\right. \\ & \left. + \log\theta_h^c\right]. \end{aligned} \tag{9}$$

Noted, λ_h is the probability, with which the current data is assigned to the hth component of combined classification criterion. For the current data, the bigger value of λ corresponds to the more suitable component.

E-step iterates Eqs. (7)–(9) till converging to the log probability of the data pair ($\boldsymbol{v}$, c). By repeating the E-step D times, we can obtain approximate posteriors of all data pair ($\boldsymbol{v}$, c). And the posteriors that will be used in M-step could facilitate the optimization of M-step.

M-step

In M-step, we maximize the lower bound on the log likelihood of the whole datasets $L(\boldsymbol{D}) = \sum_{d=1}^{D} L(\Lambda_d; \Omega)$ ($\boldsymbol{D} = \{(\boldsymbol{v}_d, c_d)|d \in \{1, 2, \ldots, D\}\}$) w.r.t model parameters $\Omega = \{\alpha, \pi, \theta^c, \eta\}$. (We do not optimize α.)

Estimating the "topic". By isolating the terms including π_{ij} and adding the appropriate Lagrange multipliers, then the terms including π_{ij} in L are

$$L_{[\pi_{ij}]} = \sum_{d=1}^{D}\sum_{m=1}^{M_d} \phi_{dm_i} \log \pi_{iv_m} + \eta_i \left(\sum_{j=1}^{V_s} \pi_{ij} - 1 \right).$$

Set $\partial L_{[\pi_{ij}]}/\partial \pi_{ij} = 0$, then

$$\pi_{ij} \propto \sum_{d=1}^{D}\sum_{m=1}^{M_d} \phi_{dmi} v_{dm}^{j}. \tag{10}$$

Estimating the "components" in combined classification criterion. The terms including η in L are:

$$L_{[\eta]} = \sum_{d=1}^{D}\sum_{h=1}^{H} \lambda_{dh} \left[\eta_{hc}^{\top} \bar{\phi}_d - \log \left(\sum_{l=1}^{C} E_q\left[\exp\left(\eta_{hl}^{\top} \bar{\boldsymbol{z}}_d \right) \right] \right) \right].$$

To compute the expectation under the variational distribution $E_q[\exp(\eta_{hl}^{\top} \bar{\boldsymbol{z}}_d)]$, we adopt the approximation which applies the multivariate delta method in [32]. The approximation is

$$\mathrm{E} f(V) \simeq f(\mathrm{E}V) + \frac{1}{2} \mathrm{tr}\left[\partial^2 f(\mathrm{E}V)/\partial v \partial v^{\top} \mathrm{cov}(V) \right], \tag{11}$$

where $f(V)$ is a mapping from R^K to R. Let $f(\bar{\boldsymbol{z}}) = \exp(\eta_{hl}^{\top} \bar{\boldsymbol{z}})$, then $\mathrm{E}_q f(\bar{\boldsymbol{z}}) \simeq \exp(\eta_{hl}^{\top} \bar{\phi})(1 + \frac{1}{2} \eta_{hl}^{\top} \mathrm{cov}(\bar{\boldsymbol{z}}) \eta_{hl})$. And $L_{[\eta]}$ is approximated as follows:

$$\begin{aligned} L_{[\eta]} \simeq \sum_{d=1}^{D}\sum_{h=1}^{H} \lambda_{dh} \Bigg[& \eta_{hc}^{\top} \bar{\phi}_d \\ & - \log \left(\sum_{l=1}^{C} \exp\left(\eta_{hl}^{\top} \bar{\phi}_d \right) (1 + \tfrac{1}{2} \eta_{hl}^{\top} \mathrm{cov}(\bar{\boldsymbol{z}}) \eta_{hl}) \right) \Bigg]. \end{aligned} \tag{12}$$

Furthermore, the derivative of $L_{[\eta]}$ with respect to η_{hc} is approximated as:

$$\begin{aligned} \frac{\partial L_{[\eta]}}{\partial \eta_{hc}} \simeq \sum_{d=1}^{D}\sum_{h=1}^{H} \lambda_{dh} \Bigg[& \bar{\phi}_d c_d^c \\ & - \frac{\exp(\eta_{hc}^{\top} \bar{\phi}_d) \left(\bar{\phi} (1 + \frac{1}{2} \eta_{hc}^{\top} \mathrm{cov}(\bar{\boldsymbol{z}}_d) \eta_{hc}) + \eta_{hc}^{\top} \mathrm{cov}(\bar{\boldsymbol{z}}_d) \right)}{\sum_{l=1}^{C} \exp(\eta_{hl}^{\top} \bar{\phi}_d)(1 + \frac{1}{2} \eta_{hl}^{\top} \mathrm{cov}(\bar{\boldsymbol{z}}_d) \eta_{hl}))} \Bigg]. \end{aligned} \tag{13}$$

where $\bar{\phi}_{di} = \sum_{m=1}^{M_d} \phi_{dmi}/M_d$ and $\mathrm{cov}(\bar{\boldsymbol{z}}_d)_{fi} = \sum_{m=1}^{M_d} (\phi_{dmi} 1(f = i) - \phi_{dmf} \phi_{dmi})/M_d^2$. It is evident that the solution has no closed form. Therefore, we adopt the conjugate gradient to tackle this problem [33]. In practice, one can add a regularization (or "weight-decay") penalty $-\xi \|\boldsymbol{\eta}\|^2$ to the objective function, as it is common for logistic regression and other classifiers.

Estimating the weights of "components". By isolating the terms including θ^c and adding the appropriate Lagrange multipliers, then we obtain the optimization objective as

$$L_{[\theta^c]} = \sum_{d=1}^{D}\sum_{h=1}^{H} \lambda_{dh} \log \theta_h^c + \eta \left(\sum_{h=1}^{H} \theta_h^c - 1 \right).$$

Maximizing the above formulation with respect to θ_h^c leads to

$$\theta_h^c \propto \sum_{d=1}^{D} \lambda_{dh}. \tag{14}$$

In summary, the variational EM algorithm alternates between E-step and M-step until the bound on the expected log likelihood converges. The procedure in detail is listed in Algorithm 1.

Algorithm 1 Variational EM for supervised latent Dirichlet allocation with a mixture of softmax.

```
Input:
  $\boldsymbol{D} = \{(\boldsymbol{v}_d, c_d)|d \in \{1, 2, \ldots, D\}\}$, the number of softmax classifiers, $H$ and thenumber of topics, $K$
Output:
  model parameters $\Omega = \{\alpha, \pi, \theta^c, \eta\}$
  repeat
    /**** E-Step ****/
    for $d = 1$ to $D$ do
      initialize $\gamma_d, \lambda_d$
      repeat
        update $\boldsymbol{\phi}_d$
        update $\gamma_d, \lambda_d$
      until $L(\Lambda_d; \Omega)$ converge
    end for
    /**** M-Step ****/
    update $\Omega = \{\alpha, \pi, \theta^c, \eta\}$
  until the log likelihood of the wholedatasets $L(\boldsymbol{D}) = \sum_{d=1}^{D} L(\Lambda_d; \Omega)$ converge
```

3.3. Supervised latent Dirichlet allocation with a mixture of sparse softmax

In MS-sLDA model, we use the L2 regularization onto the parameter η for avoiding the weight decay. Actually, a main reason we choose L2 regularization is that the derivative of L2 regularization is easily obtained. Inspired by the literature [34], in which a sparse softmax classifier with bayesian L1 regularization achieves better performance, we try to introduce the idea into a mixture of softmax, and then embed a mixture of sparse softmax into LDA model, and propose *supervised latent Dirichlet allocation with a mixture of sparse softmax*, or MSS-sLDA for short.

The MS-sLDA model and MSS-sLDA model have the same graphic model representation, generative process. So, we do not repeat them here. They also have similar parameter estimation procedures, yet the only difference is optimization of the parameter η in M-step.

In the M-step of MSS-sLDA model, the optimization of η is shown as follows. As in MS-sLDA, we isolate the terms including η in lower bound L, adopt the approximation applying the multivariate delta method in [32] and add L1 regularization of η. And then, we could obtain the optimization objective $L_{[\eta]}^{*}$ of η in MSS-sLDA model.

$$\begin{aligned} L_{[\eta]}^{*} \simeq & \sum_{d=1}^{D}\sum_{h=1}^{H} \lambda_{dh} \\ & \left[\eta_{hc}^{\top} \bar{\phi}_d - \log \left(\sum_{l=1}^{C} \exp(\eta_{hl}^{\top} \bar{\phi}_d) \left(1 + \frac{1}{2} \eta_{hl}^{\top} \mathrm{cov}(\bar{\boldsymbol{z}}) \eta_{hl} \right) \right) \right] \\ & + \mu \sum_{h=1}^{H}\sum_{l=1}^{C} |\eta_{hl}| \end{aligned} \tag{15}$$

where, $\mu > 0$ is a regularization parameter controlling the bias-variance trade-off. Based on the formulation above, we adopt the solution about derivative of L1 regularization in the literature [34], and could obtain the derivative of $L_{[\eta]}$ with respect to η_{hc} as:

$$\frac{\partial L^{*}_{[\eta]}}{\partial \eta_{hc}} = \begin{cases} \frac{\partial L_{[\eta]}}{\partial \eta_{hc}} + \mu, & \eta_{hc} > 0 \\ \frac{\partial L_{[\eta]}}{\partial \eta_{hc}} - \mu, & \eta_{hc} < 0 \\ \frac{\partial L_{[\eta]}}{\partial \eta_{hc}} + \mu, & \eta_{hc} = 0 \, and \, \frac{\partial L_{[\eta]}}{\partial \eta_{hc}} + \mu < 0 \\ \frac{\partial L_{[\eta]}}{\partial \eta_{hc}} - \mu, & \eta_{hc} = 0 \, and \, \frac{\partial L_{[\eta]}}{\partial \eta_{hc}} + \mu > 0 \\ 0, & otherwise \end{cases} \tag{16}$$

where $\frac{\partial L_{[\eta]}}{\partial \eta_{hc}}$ is the same as in the MS-sLDA model, and we could consult Eq. (14) for the details of $\frac{\partial L_{[\eta]}}{\partial \eta_{hc}}$.

3.4. Predicting category

In the previous sections, we have introduced the proposed models and the procedure of estimating model parameters. In this section, we describe how to use the models to predict class label of unseen data. In particular, we use latent topic frequencies $\bar{z}$ of a data instead of original low-level features. The posterior probability over latent topics, however, is computationally intractable. So, we use $\mathrm{E}_q[\bar{z}] = \bar{\phi}$, which can be computed using the variational E-step of LDA [9], to approximate $\bar{z}$, and $\bar{\phi}$ could be obtained by removing the terms including λ from Eq. (6) and the terms including η and λ from Eq. (8).

After obtaining approximate latent topic frequencies as well as all softmax classifiers, a direct idea is to adopt the classical SMM method to predict the class label of a data. In the predicting method of SMM, however, the final output is the weighted average of the outputs obtained by all softmax classifiers, so that the poor output of some a softmax classifier with bigger weights will result in inaccurate decision. Therefore, in order to reduce this kind of predicting errors, the weighted average is abandoned. Instead, the biggest value is selected directly from the outputs of all softmax, and the class label corresponding to the biggest value will be assigned to the data. In particular, the formulation is

$$\begin{aligned} C_{*} &= \arg \max_{h\in\{1,2,\ldots,H\},c\in\{1,2,\ldots,C\}} \frac{\exp(\eta_{hc}^{\top}\bar{z})}{\sum_{l=1}^{C}\exp(\eta_{hl}^{\top}\bar{z})} \\ &\simeq \arg \max_{h\in\{1,2,\ldots,H\},c\in\{1,2,\ldots,C\}} \frac{\exp(\eta_{hc}^{\top}\bar{\phi})}{\sum_{l=1}^{C}\exp(\eta_{hl}^{\top}\bar{\phi})}. \end{aligned} \tag{17}$$

From the formulation, our ensemble method is the “disjunction” of all softmax classifiers. Therefore, it could take the advantage of every softmax classifier and obtain a better combined classifier.

4. Experiments

We choose the following three datasets:

- Scene classification dataset: a subset of LabelMe dataset from [11]. The LabelMe data contains natural scene images with 8 classes. We randomly selected 200 images for each class with the total number of images 1600.
- Event classification datasets: UIUC-Sport data from [7]. The UIUC-Sport data contains sports scene images with 8 classes. The number of images in each class varies from 137 (bocce) to 329 (croquet), and the total number of images is 1791. See the Table 1 for the details.
- Text classification dataset: 20 Newsgroups dataset. The dataset contains about 18774 postings in 20 related categories, and the number of each class is listed in Table 2.

Table 1
UIUC-Sport dataset.

Classes	Sample size
Badminton	313
Bocce	137
Croquet	329
Polo	183
Rockclimbing	194
Rowing	255
Sailing	190
Snowboarding	190

Table 2
20 Newsgroups dataset.

Classes	Sample size
Alt.atheism	798
Comp.graphics	970
Comp.os.ms-windows.misc	963
Comp.sys.ibm.pc.hardware	979
Comp.sys.mac.hardware	958
Comp.windows.x	982
Misc.forsale	964
Rec.autos	987
Rec.motorcycles	993
Rec.sport.baseball	991
Rec.sport.hockey	997
Sci.crypt	989
Sci.electronics	984
Sci.med	987
Sci.space	985
Soc.religion.christian	997
Talk.politics.guns	909
Talk.politics.mideast	940
Talk.politics.misc	774
Talk.religion.misc	627

Since the two proposed models are built on “bag-of-words” representation, the preprocessing of data is transforming an original data into a vector of word counts. For LabelMe data, the preprocessing steps are: 1. Extract 16 × 16-size patch applying grid sampling (5 × 5 grid size) technique for all images, then use 128-dimensional SIFT [35] region descriptor to represent each patch. 2. Run the k-means algorithm [36] for the descriptors collections, and construct the codebook using all centers obtained by k-means. (The codebook size is set as 240.) 3. Label the patches in each image using the code words. Finally, each image is represented as a vector of codeword counts.

For UIUC-Sport data, we adopt similar steps as those for LabelMe dataset, and the only difference is that we extract 2500 patches uniformly for every image in this dataset with the size of each patch 32 × 32.

For 20 Newsgroups data, we use the training/testing split in webpage http://people- .csail.mit.edu/jrennie/20Newsgroups/. And the training data contains 11,269 documents and the test data contains 7505 documents.

4.1. Classification accuracy

In order to evaluate the classification performance of the two proposed models, we compare them with the following methods: (1) *Multi-class sLDA* (MC-sLDA) [17], (2) *Supervised Document Neural Autoregressive Distribution Estimator model* (Sup-NADE) [24], (3) *maximum entropy discrimination latent Dirichlet allocation model*

Table 3
Comparisons of average accuracy over all classes on the UIUC-Sport dataset based on 5 random train/test subsets: MC: *MC-sLDA*. MS-*: the number of softmax in MS-sLDA model is set as*. MSS-*: the number of softmax in MSS-sLDA model is set as*.

Topics	MC	MS-2	MS-3	MS-4	MS-5	MS-6	MSS-2	MSS-3	MSS-4	MSS-5	MSS-6
20	0.546	0.653	0.631	0.631	0.620	0.601	0.650	0.671	0.680	0.690	0.650
40	0.615	0.681	0.680	0.684	0.654	0.620	0.670	0.690	0.697	0.701	0.680
60	0.632	0.661	0.668	0.691	0. 678	0.640	0.684	0.692	0.720	0.719	0.716
80	0.634	0.665	0.703	0.695	0.701	0.689	0.680	0.687	0.732	0.729	0.725
100	0.635	0.710	0.728	***0.733***	0.732	0.710	0.721	0.718	***0.744***	0.740	0.738
120	***0.640***	0.699	0.731	0.697	0.710	0.727	0.721	0.740	0.720	0.719	0.719

Table 4
Comparisons of average accuracy over all classes on the LabelMe dataset based on 5 random train/test subsets: MC: *MC-sLDA*. MS-*: the number of softmax in MS-sLDA model is set as*. MSS-*: the number of softmax in MSS-sLDA model is set as*.

Topics	MC	MS-2	MS-3	MS-4	MS-5	MS-6	MSS-2	MSS-3	MSS-4	MSS-5	MSS-6
20	0.724	0.795	0.783	0.777	0.767	0.746	0.791	0.820	0.791	0.781	0.780
40	0.748	0.819	0.795	0.799	0.787	0.770	0.811	0.831	0.829	0.810	0.797
60	0.761	0.817	0.816	0.824	0.810	0.786	0.791	0.832	0.840	0.830	0.810
80	***0.766***	0.825	0.831	***0.841***	0.836	0.799	0.831	***0.860***	0.843	0.844	0.801
100	0.755	0.836	0.827	0.838	0.827	0.800	0.826	0.832	0.841	0.832	0.821
120	0.758	0.831	0.820	0.825	0.810	0.802	0.834	0.835	0.838	0.832	0.809

Table 5
Comparisons of average accuracy over all classes on the 20 Newsgroups dataset based on 5 random train/test subsets: MC: *MC-sLDA*. MS-*: the number of softmax in MS-sLDA model is set as*. MSS-*: the number of softmax in MSS-sLDA model is set as*.

Topics	MC	MS-2	MS-3	MS-4	MS-5	MS-6	MSS-2	MSS-3	MSS-4	MSS-5	MSS-6
20	0.720	0.781	0.783	0.777	0.767	0.760	0.791	0.786	0.760	0.762	0.765
40	0.750	0.794	0.795	0.799	0.780	0.770	0.799	0.812	0.814	0.810	0.803
60	0.762	0.798	0.801	0.804	0.798	0.787	0.800	0.806	0.810	0.809	0.802
80	0.690	0.786	***0.810***	0.801	0.800	0.792	0.810	0.814	0.806	0.802	0.799
100	***0.764***	0.800	0.806	0.798	0.787	0.772	***0.817***	0.810	0.811	0.814	0.808
120	0.732	0.803	0.809	0.799	0.784	0.769	0.799	0.801	0.810	0.802	0.798

(Med-LDA) [18]. (4) *Sparse Bayesian Multinomial Logistic Regression* (SBMLR) [34]), (5)*SVM with polynomial kernel* (SVM-POL).

In these methods, MC-sLDA and Sup-NADE and Med-sLDA are supervised topic model for classification. SBMLR, a softmax function with Laplace prior, shows good performance on many datasets. SVM-POL are widely used and proved to be effective classification methods. (We use libsvm package [37] in our experiment.)

Tables 3–5 show the performance of the two proposed models and MC-sLDA on the three datasets. The first column of each table labels dataset and number of topics, the second column reports accuracies of MC-sLDA, and the remaining columns show accuracies of the two proposed models when $H = *$ is set as different numbers.

From Table 3, on UIUC-Sport data, MS-sLDA (H=4, K=100) has the best performance 73.3%, which is 9% higher than MC-sLDA ($K = 120$), and MSS-sLDA (H= 4, $K = 100$) has the best performance 74.4% which is 10% higher than MC-sLDA (K= 120). And from Table 4, it can be seen that: on LabelMe data, MS-sLDA ($H = 4$, $K = 80$) has the best performance 84.1%, and is about 7% higher than MC-sLDA ($K = 80$). While MSS-sLDA ($H = 3$, $K = 80$) has the best performance 86.0%, and is about 9% higher than MC-sLDA ($K = 80$). Moreover, from Table 5, MS-sLDA ($H = 3$, K= 80) has the best performance 81.0% on 20 Newsgroups dataset, and is about 4.6% higher than MC-sLDA ($K = 100$). It also can be seen that MSS-sLDA ($H = 2$, $K = 100$) has the best performance 81.0%, and is about 5.4% higher than MC-sLDA ($K = 100$).

We also test Sup-NADE, Med-sLDA, SBMLR and SVM-POL on the three datasets, see Table 6. In Table 6, the first two columns show the performance of SVM-POL and SBMLR, the following three columns list the best performance of MC-sLDA, Sup-NADE and Med-sLDA model. The last two columns report the best performance of the proposed MS-sLDA and MSS-sLDA model. And the obtained average accuracy of SBMLR is 64.3% for UIUC-Sport data, 74.3% for LabelMe data, and 79.8% for 20 Newsgroups data, respectively. The average accuracy of SVM-POL is 52% for UIUC-Sport data, 69.5% for LabelMe data, and 80.3% for 20 Newsgroups data, respectively. On the three datasets, three topic models, Sup-NADE, MC-sLDA and Med-sLDA, perform better than SVM-POL and SBMLR, and the proposed models perform better than these three topic models.

In sum, the results show the two proposed models achieve a superior performance comparing with other benchmark methods.

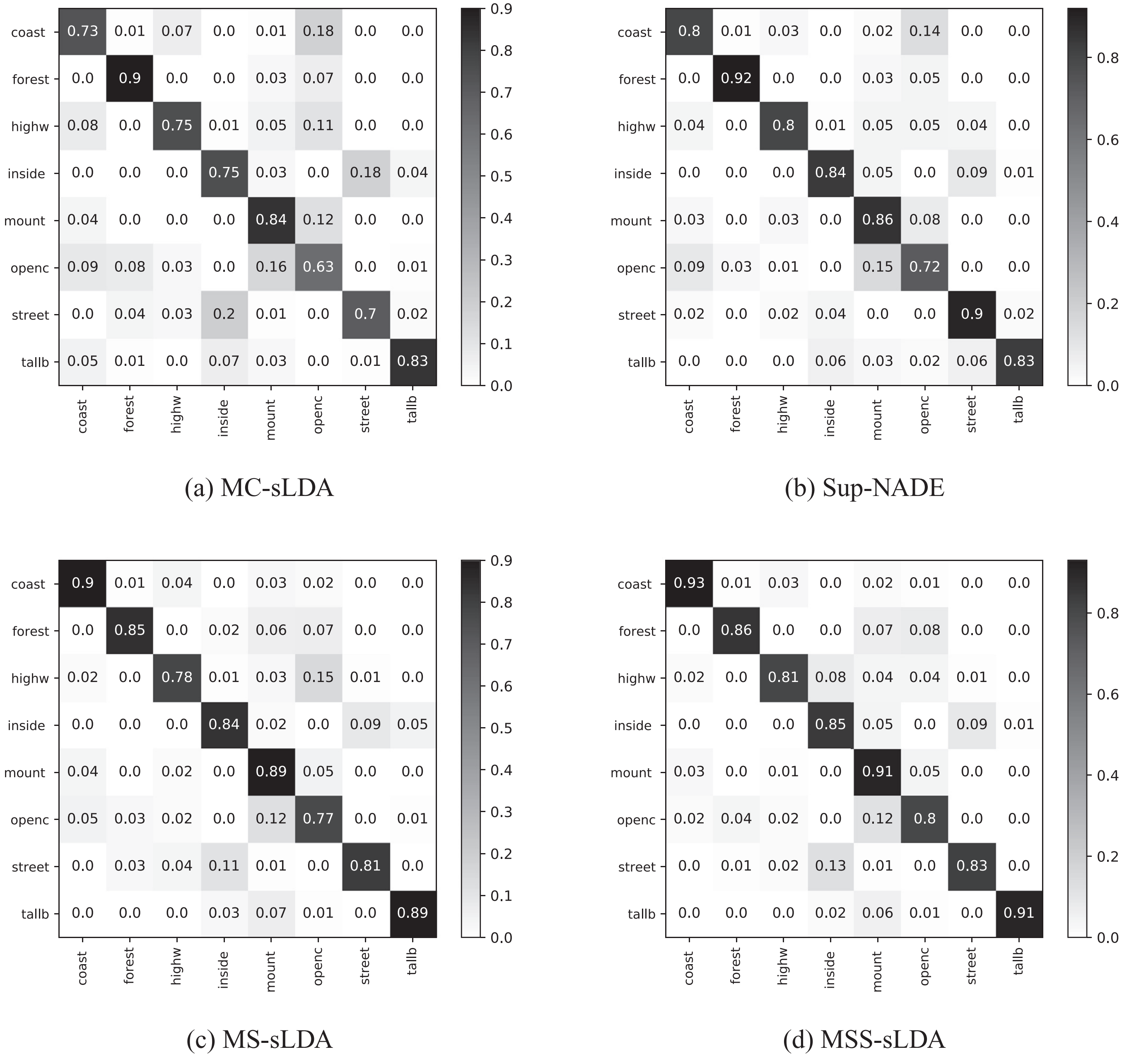
4.2. Confusion matrix

To further evaluate classification performance the two proposed models, we show confusion matrices in Fig. 5. Because of the space, we only show confusion matrix of the best performance of Mc-sLDA, sup-NADE, MS-sLDA and MSS-sLDA on LabelMe Dataset. In particular, MC-sLDA is with 80 topics, sup-NADE is with 120 topics, MS-sLDA is with 80 topics and 4 softmax classifiers, and MSS-sLDA is with 80 topics and 4 softmax classifiers.

From Fig. 5, we can see that classifying LabelMe data by using MC-sLDA and sup-NADE, “coast” and “opening country” class have big confusion. It is easily to classify a “coast” image into “opening country” class, and also easily to classify a “opening country” image into “coast” class. That means MC-sLDA and sup-NADE model are difficult to classify these two classes, while the proposed models, MS-sLDA and MSS-sLDA, could improve the situation.

Table 6
Comparisons of average accuracy over all classes based on 5 random train/test subsets. UIUC: *UIUC-sport data*. LabelMe: *the subset of LabelMe datasets*. 20-News: *20 Newsgroups datasets*. Sup-NADE: Supervised Document Neural Autoregressive Distribution Estimator model. Med-LDA: maximum entropy discrimination latent Dirichlet allocation model. MS-sLDA and MSS-sLDA are the two proposed models.

DataSets	SVM-POL	SBMLR	MC-sLDA	Sup-NADE	Med-sLDA	MS-sLDA	MSS-sLDA
UIUC	0.520	0.643	0.640	0.69	0.67	***0.733***	***0.744***
LabelMe	0.695	0.748	0.766	0.820	0.780	***0.841***	***0.860***
20-News	0.803	0.798	0.764	0.772	0.800	***0.810***	***0.817***



(a) MC-sLDA

(b) Sup-NADE

(c) MS-sLDA

(d) MSS-sLDA

Fig. 5. Comparisons using confusion matrices on LabelMe dataset, the confusion matrices from the best performance of MC-sLDA (80 topics), Sup-NADE (120 topics), MS-sLDA(80 topics), and MSS-sLDA (80 topics).

So, our models are effective for classifying the kind of data in which within-class similarity is relatively low, while the between-class similarity is relatively high.

4.3. “Components” in combined classification criterion

This part will visualize classification criterion of MS-sLDA model with a mixture of 2 softmax classifiers on UIUC-Sport dataset. From Table 3, the average accuracy of MS-sLDA ($H = 2$, $k = 20$) on UIUC-Sport data is 65.3%. It reduces the error of MC-sLDA ($K = 20$) 9.7%, and also outperforms the best performance of MC-sLDA. In Fig. 6, the two gray-scale images on the left side represent two classification “components” (softmax classifier) learned by our model (H= 2, k = 20). The “components” can be obtained by transforming the parameters (a 8×20 matrix, and 8 and 20 represents the number of classes and topics, respectively)

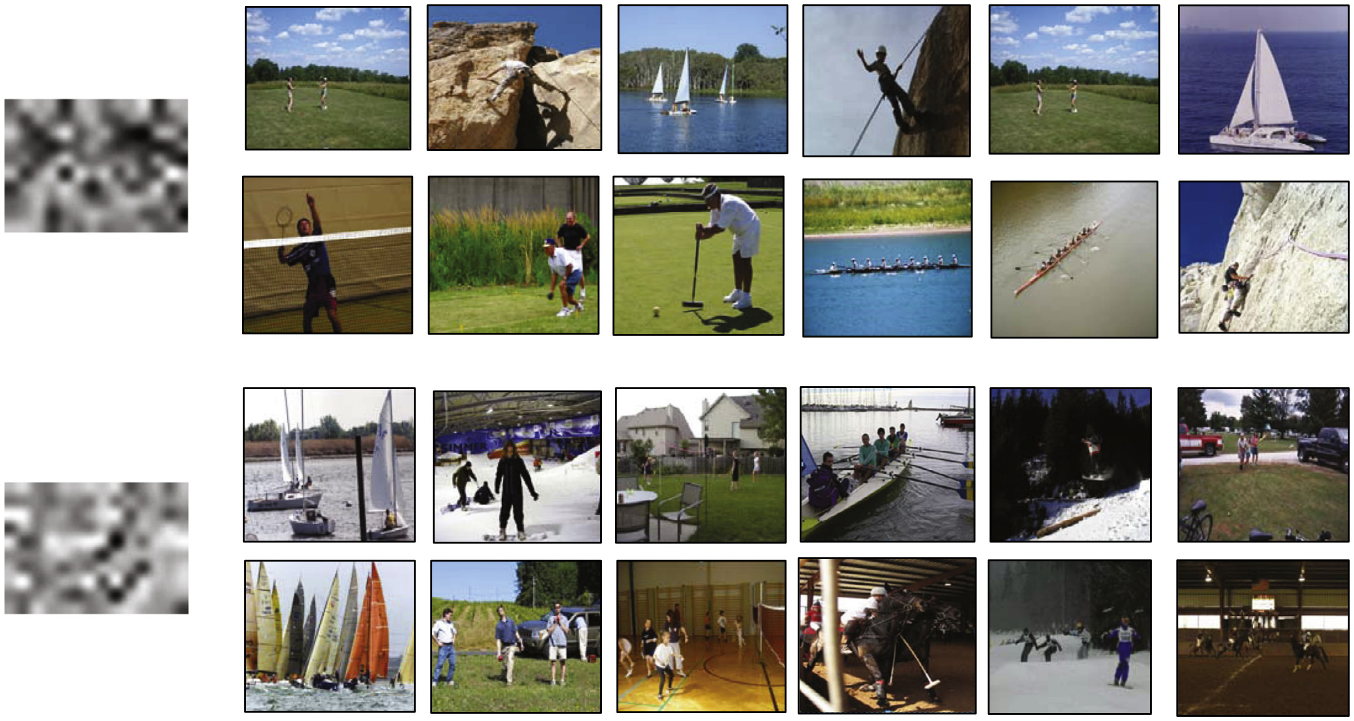


Fig. 6. The different components in combined classification criterion (2 grayscale images on the left side) learned by Mixture of softmax sLDA model ($H = 2$, $K = 20$) on UIUC-sport dataset, as well as example images (16 images on the right side) assigned to the corresponding component (softmax classifier).

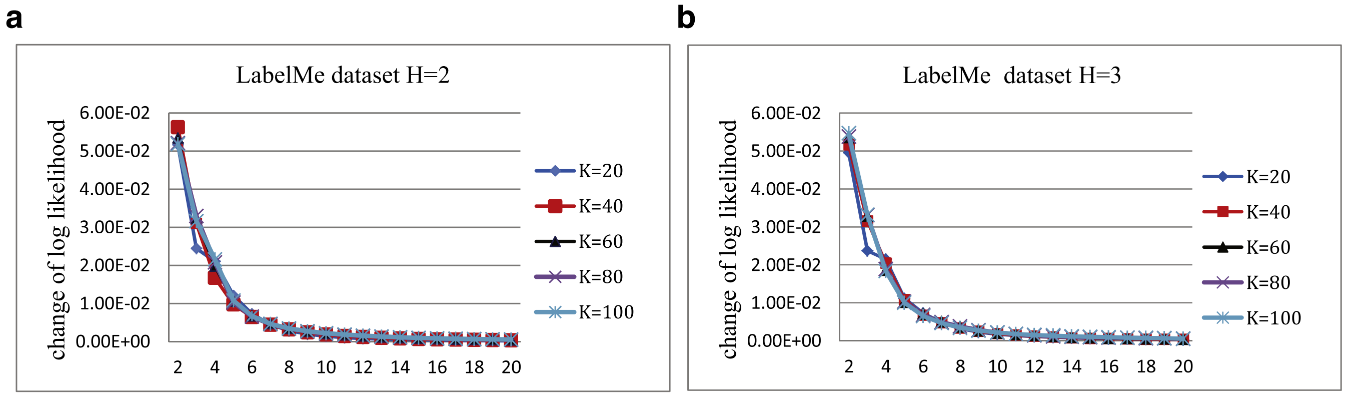


Fig. 7. The vertical axis records the relative change of log-likelihood function of MS-sLDA model on the LabelMe dataset, and the horizontal axis records the number of iterations EM algorithm. K represents topic numbers. (a). The number of classifiers is set as 2. (b). The number of classifiers is set as 3.

of each softmax classifier to a gray-scale image, in which the more brighter pixel means the larger value. And the real-scene images on the right side are some examples, which are assigned to the corresponding “component”. In E-step, the variational parameter λ_h records the component assignment of the current image. For the hth classification component in the combined classifier, we select some example images with bigger value of λ_h. From Fig. 6, it is easy to find the two components are different. One is with the darker, its example images contain fewer objects. And the other is brighter, its example images contain more objects. It further shows MS-sLDA ($H = 2$, $k = 20$) can learn multiple classification components, and each component can be used to fit a special type of image subsets.

MS-sLDA and MSS-sLDA model with different topic numbers and classifier numbers also show the similar situation, so we do not repeat showing the details here.

4.4. Discussion

According to the experimental results, the two proposed models show better performance than MC-sLDA and other baseline methods. The reason is that the two proposed models introducing ensemble classification idea to sLDA model, learn latent semantic feature and multiple criteria for classification simultaneously, and combine the advantages of multiple criteria. In addition, in the two proposed models, MSS-sLDA model achieves better performance than MS-sLDA model. The season is mainly that MSS-sLDA model adopts a kind of effective L1 regularization method, which learns from the literature [34], for ensemble classification part in the model, while MS-sLDA model adopts L2 regularization in the corresponding part.

Although two proposed models have better performance, two proposed models have more difficulties on model selection. From

Tables 3–5, we find that on these three datasets, if we fix topic number (such as 60) and increase the number of softmax classifiers, such as from 2 to 6, the performance goes up first, and then goes down. 3 or 4 are the best number of softmax classifiers on these datasets, 5 and 6 could make our models overfit. If we fix the number of softmax classifiers and increase topic number, such as from 20 to 120, the situation is also similar. The hyperparameters, the number of softmax classifiers and the number of topics, are key factor of effecting the performance of the two proposed models.

In general, models of LDA series optimized by EM algorithm are fit for short text data. Compared to other methods reducing dimension [38], this type of models has no advantage on training time. The two proposed models have more parameters than classical LDA model, that means they need more training time. Especially, for lengthy and books dataset in which the dimension of a data point would be extremely high since a book usually covers the whole vocabulary, such as the one used in [39], our proposed models will have very prohibitive time cost and hardware cost. So, our models will be more suitable for small sample data in which the dimensions of data points are not extremely high.

5. Analysis of convergence

Till now, we have introduced the proposed model, *MS-sLDA* and *MSS-sLDA*, and their the parameter estimation procedure. In this section, we only take MS-sLDA model for example and discuss the convergence of our parameter estimation procedure. For convergence analysis of MSS-sLDA, it is similar to MS-sLDA.

The procedure is derived from the variational EM framework, in which the lower bound of log-likelihood ascends coordinately in E-step and M-step. Because log-likelihood has upper bound, if both E-step and M-step can hold the lower bound, then the procedure can converge. In E-step, we maximize the lower bound $L(\Lambda_d; \Omega)$ w.r.t. variational parameters Λ_d for arbitrary $d \in 1, \ldots, D$, and do not make approximation. So the step holds the bound. In M-step, however, we maximize the lower bound $L(\Lambda_d; \Omega)$ w.r.t. model parameters $\Omega = \{\alpha, \pi, \theta^c, \eta\}$, and optimization of η in M-step can not hold the lower bound at any condition.

Theorem 1. *For arbitrary $d \in 1, \ldots, D$, if the maximum of $L(\Lambda_d; \Omega)$ w.r.t. variational parameters Λ_d can be found after E-step for arbitrary $d \in 1, \ldots, D$, and M_d is large enough, then optimization of η can hold the lower bound.*

Proof. In optimization of η, we make two approximations. The first approximation adopting Jensen's inequality holds the lower bound. However, the second approximation adopting Eq. (11) can not hold the lower bound.

For arbitrary $d \in 1, \ldots, D$, if the maximum of $L(\Lambda_d; \Omega)$ w.s.t.variational parameters Λ_d can be found after E-step, then we can know $KL(p(z|E, \theta) \| q(z|\phi)) \rightarrow 0$ after E-step. Namely, $\mathrm{E}_q[\bar{z}_d] = \bar{\phi}_d$ tends to posterior mean $\mathrm{E}_{p(z|E, \theta)} z_d$. And according to law of large numbers, if M_d is enough large, then empirical mean $\bar{z}_d = \frac{\sum_{m=1}^{M_d} z_{dm}}{M_d}$ tends to posterior mean $\mathrm{E}_{p(z|E, \theta)} z$. So, $\bar{z}_d$ approaches $\mathrm{E}_q[\bar{z}_d] = \bar{\phi}_d$.

And because when $\mu \rightarrow \mathrm{E}\mu$, the left hand of Eq. (11) tends to the right hand,

$$\mathrm{E} f(\mu) \rightarrow f(\mathrm{E}\mu) + \frac{1}{2} \mathrm{tr}\left[\partial^2 f(\mathrm{E}\mu) / \partial \mu \partial \mu^\top \mathrm{cov}(\mu)\right]$$

Namely,

$$\mathrm{E}_q[\exp(\eta_l^\top \bar{z}_d)] \rightarrow \exp(\eta_l^\top \bar{\phi}_d)\left(1 + \frac{1}{2} \eta_l^\top \mathrm{cov}(\bar{z}_d) \eta_l\right)$$

This means for arbitrary $d \in 1, \ldots, D$, if the maximum of $L(\Lambda_d; \Omega)$ w.s.t. variational parameters Λ_d can be found after E-step, and M_d is enough large, the approximation Eq. (11) can hold the lower bound. That is, optimization of η can hold the lower bound. □

To sum up, when the conditions in Theorem 1 can be satisfied, M-step can hold the lower bound. And because the lower bound ascends coordinately in E-step and M-step, and log-likelihood has upper bound, so MS-sLDA model can converge when the conditions in Theorem 1 can be satisfied.

Actually, the conditions are so rigorous that they cannot be met in many cases. In the experiments above, the lengths of data M_d in three datasets are from several dozens to several hundreds. Obviously, the condition is not met in our experiments. Therefore, as shown in the next section, we use the change of likelihood to examine the convergence of MS-sLDA model.

5.1. The change of likelihood

The relative change of log-likelihood P is defined as follows

$$P_{t+1} = \frac{\mathrm{loglikelihood}_{t+1} - \mathrm{loglikelihood}_t}{\mathrm{loglikelihood}_t}, \tag{18}$$

in which t represents the iteration number of EM algorithm. The termination condition of the parameter estimate procedure in MS-sLDA model is $P_{t+1} < \varepsilon$, where ε is a constant which is set in advance. (In our experiment, $\varepsilon = 1e - 4$).

Fig. 7 shows the change of P_t on LabelMe dataset when MS-sLDA model parameter is set to different values. The vertical axis records Pt and horizontal axis records the number of iterations EM algorithm in our parameter estimation procedure. K represents the topic numbers. In Fig. 7, when topic is set as different numbers, P_{t+1} always become smaller than P_t as the number of iterations increases in two datasets. And these two values P_t and P_{t+1} get close to zero as time goes. It shows MS-sLDA model can converge on LabelMe dataset. We also compute the change of likelihood on another two datasets, and get similar results.

In other words, even though the convergence conditions in Theorem 1 have not been met, our parameter estimation procedure could still converge.

6. Conclusion and future work

6.1. Conclusion

In the paper, we proposed two supervised topic models with ensemble classification merits, namely *MS-sLDA* and *MSS-sLDA*. We given the procedure of parameter estimation, presented an efficient approximation method for predicting the class label of an unseen data, and demonstrated the effectiveness of the two proposed models on the real datasets. Experimental results show the two proposed models can achieve better performance. From experimental results and classification criteria visualization, we can conclude that a single criterion is difficult to fit the relationship between data and label, and a combined criterion is more reasonable. Finally, we conducted a theoretical and experimental analysis for the convergence of the proposed models.

The proposed models have the characters of learning latent semantic in topic model and ensemble classification. In the learning phase, it constructs multiple classification criteria to fit the training data. In the test phase, it combines the advantage of all classifiers to construct a strong combined classifier. These proposed models belong to supervised topic model fusing ensemble concept, and are different from conventional ensemble classification methods, such as [40]. In [40], two independent classifiers for visual features and text features based on SVM-POL are constructed firstly, and then a third logistic regression classifier is trained to combine the confidence values of the two initial classifiers into a final prediction. The significant difference is that the two SVM classifiers in

[40] fit visual feature and text feature respectively. In MS-sLDA and MSS-sLDA models, however, multiple components (classifiers) are correlated during training process, each component is merely used to fit a subset of training dataset.

6.2. Future work

Although the proposed models show better classification performance, there are many points worth further improvement and in-depth study in the future. Firstly, the proposed models have two hyper-parameters, the number of topics and the number of softmax classifiers. To reduce the burden of model selection, we will investigate how to learn the parameters adaptively, which also could increase the difficulties of solving parameters. Secondly, in our model, the relationship between data and classifiers is that given a data point, it could choose a classifier which has higher prediction probability on the data point under the condition of maximizing the likelyhood of training data. We use this kind of relationship to achieve that different data points could choose different classifiers. We will try that let the selection of classifiers be dependent on empirical topic frequencies, and do further improvement in the future work. Finally, the proposed models provided a method of combing ensemble ideas and supervised topic models. Naturally, how to design a general framework of it is also valuable to be investigated.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China (NSFC) grant nos. 61563030, 61773071, the Natural Science Foundation of Gansu Province, China, grant no. 2017GS10830, Beijing Natural Science Foundation (BNSF) grant no. 4162044, Beijing Nova Program grant no. Z171100001117049, CCF-Tencent Open Research Fund.

References

[1] Z. Ma, A. Leijon, Bayesian estimation of beta mixture models with variational inference, IEEE Trans. Pattern Anal. Mach. Intell. 33 (11) (2011) 2160–2173.

[2] C. Bishop, Pattern Recognition and Machine Learning, 4, Springer, New York, 2006.

[3] G. Xun, Y. Li, W.X. Zhao, J. Gao, A. Zhang, A correlated topic model using word embeddings, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017, pp. 4207–4213.

[4] M. Rabinovich, D. Blei, The inverse regression topic model, in: Proceedings of the International Conference on Machine Learning, 2014, pp. 199–207.

[5] L. Fei-Fei, P. Perona, A Bayesian hierarchical model for learning natural scene categories, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2, IEEE, 2005, pp. 524–531.

[6] F. Xue, J. Wang, S. Qian, T. Zhang, X. Liu, C. Xu, Multi-modal max-margin supervised topic model for social event analysis, Multimed. Tools Appl. 77 (2018) 1–20.

[7] L. Li, L. Fei-Fei, What, where and who? classifying events by scene and object recognition, in: Proceedings of the IEEE Eleventh International Conference on Computer Vision, IEEE, 2007, pp. 1–8.

[8] D. Blei, J. McAuliffe, Supervised topic models. Proceedings of the Advances in Neural Information Processing Systems, 2008, 121–128.

[9] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[10] M. Jordan, Z. Ghahramani, T. Jaakkola, L. Saul, An introduction to variational methods for graphical models, Mach. Learn. 37 (2) (1999) 183–233.

[11] B. Russell, A. Torralba, K. Murphy, W. Freeman, Labelme: a database and web-based tool for image annotation, Int. J. Comput. Vis. 77 (1) (2008) 157–173.

[12] R. Das, M. Zaheer, C. Dyer, Gaussian LDA for topic models with word embeddings, in: Proceedings of the Fifty-Third Annual Meeting of the Association for Computational Linguistics and the Seventh International Joint Conference on Natural Language Processing, 1, 2015, pp. 795–804.

[13] K.W. Lim, C. Chen, W. Buntine, Twitter-network topic model: a full Bayesian treatment for social network and text modeling, (2016). arXiv:1609.06791.

[14] S. Qian, T. Zhang, C. Xu, J. Shao, Multi-modal event topic model for social event analysis, IEEE Trans. Multimed. 18 (2) (2016) 233–246.

[15] N. Chen, J. Zhu, F. Sun, B. Zhang, Learning harmonium models with infinite latent features, IEEE Trans. Neural Netw. Learn. Syst. 25 (3) (2014) 520–532.

[16] H. Zhang, Y. Ji, J. Li, Y. Ye, A triple wing harmonium model for movie recommendation, IEEE Trans. Ind. Inf. 12 (1) (2016) 231–239.

[17] C. Wang, D. Blei, F. Li, Simultaneous image classification and annotation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 1903–1910.

[18] J. Zhu, L.-J. Li, L. Fei-Fei, E.P. Xing, Large margin learning of upstream scene understanding models, in: Proceedings of the Advances in Neural Information Processing Systems, 2010, pp. 2586–2594.

[19] I. Titov, R. McDonald, A joint model of text and aspect ratings for sentiment summarization, Urbana 51 (2008) 61801.

[20] L.-J. Li, R. Socher, L. Fei-Fei, Towards total scene understanding: classification, annotation and segmentation in an automatic framework, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 2036–2043.

[21] S. Lacoste-Julien, F. Sha, M.I. Jordan, DiscLDA: discriminative learning for dimensionality reduction and classification, in: Proceedings of the Advances in Neural Information Processing Systems, 2008, pp. 897–904.

[22] J. Zhu, A. Ahmed, E.P. Xing, MedLDA: maximum margin supervised topic models, J. Mach. Learn. Res. 13 (1) (2012) 2237–2278.

[23] Y. Wang, G. Mori, Max-margin latent Dirichlet allocation for image classification and annotation., in: Proceedings of the Twenty-Second British Machine Vision Conference, 2011, pp. 1–11.

[24] Y. Zheng, Y.-J. Zhang, H. Larochelle, Topic modeling of multimodal data: an autoregressive approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2014, pp. 1370–1377.

[25] Z. Si, R. Thobaben, M. Skoglund, Rate-compatible LDPC convolutional codes achieving the capacity of the BEC, IEEE Trans. Inf. Theory 58 (6) (2012) 4021–4029.

[26] Z. Si, R. Thobaben, M. Skoglund, Bilayer LDPC convolutional codes for decode-and-forward relaying, IEEE Trans. Commun. 61 (8) (2013) 3086–3099.

[27] X. Li, J. Zeng, X. Wang, Y. Zhong, Mixture of softmax slLDA, in: Proceedings of the Eleventh IEEE International Conference on Data Mining (ICDM), IEEE, 2011, pp. 1164–1169.

[28] Z. Ma, J.-H. Xue, A. Leijon, Z.-H. Tan, Z. Yang, J. Guo, Decorrelation of neutral vector variables: theory and applications, IEEE Trans. Neural Netw. Learn. Syst. 29 (1) (2018) 129–143.

[29] W. Chen, Simultaneous sparse Bayesian learning with partially shared supports, IEEE Signal Process. Lett. 24 (11) (2017) 1641–1645.

[30] Z. Ma, A.E. Teschendorff, A. Leijon, Y. Qiao, H. Zhang, J. Guo, Variational Bayesian matrix factorization for bounded support data, IEEE Trans. Pattern Anal. Mach. Intell. 37 (4) (2015) 876–889.

[31] W. Chen, D. Wipf, Y. Wang, Y. Liu, I.J. Wassell, Simultaneous Bayesian sparse approximation with structured sparse models, IEEE Trans. Signal Process. 64 (23) (2016) 6145–6159.

[32] M. Braun, J. McAuliffe, Variational inference for large-scale models of discrete choice, J. Am. Stat. Assoc. 105 (489) (2010) 324–335.

[33] J. Nocedal, S. Wright, Numerical Optimization, Springer verlag, 1999.

[34] G. Cawley, N. Talbot, M. Girolami, Sparse multinomial logistic regression via Bayesian L1 regularisation, Adv. Neural Inf. Process. Syst. 19 (2007) 209.

[35] D. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the Seventh IEEE International Conference on Computer Vision, 2, IEEE, 1999, pp. 1150–1157.

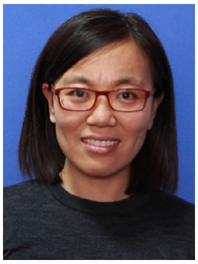
[36] T. Kadir, M. Brady, Saliency, scale and image description, Int. J. Comput. Vis. 45 (2) (2001) 83–105.

[37] C. Chang, C. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. (TIST) 2 (3) (2011) 27.

[38] Z. Ma, J. Xie, H. Li, Q. Sun, Z. Si, J. Zhang, J. Guo, The role of data analysis in the development of intelligent energy networks, IEEE Netw. 31 (5) (2017) 88–95.

[39] H. Zhang, T.W. Chow, Q.J. Wu, Organizing books and authors by multilayer SOM, IEEE Trans. Neural Netw. Learn. Syst. 27 (12) (2016) 2537–2550.

[40] G. Wang, D. Hoiem, D. Forsyth, Building text features for object image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 1367–1374.



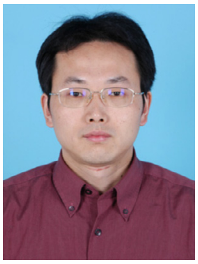
Xiaoxu Li received his Ph.D. degree from Beijing University of Posts and Telecommunications (BUPT), China, in 2012. She is an associate professor at Lanzhou University of Technology in the School of computer and communication. Her research interests include machine learning and computer vision. She is a member of China Computer Federation.

Zhanyu Ma has been an associate Professor at Beijing University of Posts and Telecommunications (BUPT), Beijing, China, since 2014. He received his Ph.D. degree in Electrical Engineering from KTH (Royal Institute of Technology), Sweden, in 2011. From 2012–2013, he has been a Postdoctoral research fellow in the School of Electrical Engineering, KTH, Sweden. His research interests include statistical modeling and machine learning related topics with a focus on applications in speech processing, image processing.

Pai Peng received the Ph.D. degree in computer science from Zhejiang University in 2016. He is currently a research scientist in Youtu Lab of Tencent Technology (Shanghai) Co., Ltd. His research interests include image recognition and deep learning and has published several top-tier conference and journal papers related with image recognition, e.g. SIGIR, CIKM, TKDE, ICMR, etc.

XiaoWei Guo received his M.S. degree in Information and Computing Science from Sun Yat-sen University in 2007 and now works as a senior engineer in YouTu lab of Tencent Technology (Shanghai) Co., Ltd. He is responsible for the R & D and management of the Image Understanding team. His research interests include Image Recognition, OCR, Deep Learning, Transfer Learning, Augmented Reality / Virtual Reality.

Feiyue Huang received his B.Sc. and Ph.D. degrees in Computer Science in 2001 and 2008, both from Tsinghua University, China. He is the director of Tencent Youtu Lab. His research interests include machine learning and computer vision.



Xiaojie Wang received his Ph.D. degree from Beihang University in 1996. He is a professor and director of the Centre for Intelligence Science and Technology at Beijing University of Posts and Telecommunications. His research interests include natural language processing and multimodal cognitive computing. He is an executive member of the Council of Chinese Association of Artificial Intelligence, director of Natural Language Processing Committee. He is a member of Council of Chinese Information Processing Society and Chinese Processing Committee of China Computer Federation.

Jun Guo received B.E. and M.E. degrees from Beijing University of Posts and Telecommunications (BUPT), China in 1982 and 1985, respectively, Ph.D. degree from the Tohuku-Gakuin University, Japan in 1993. At present he is a professor and a vice president of BUPT. His research interests include pattern recognition theory and application, information retrieval, content based information security, and network management.