Neurocomputing 134 (2014) 38–45

Contents lists available at ScienceDirect

Neurocomputing

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

A jointly distributed semi-supervised topic model

Yanning Zhang, Wei Wei *

Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, School of Computer Science, Northwestern Polytechnical University, Xi'an, China

ARTICLE INFO

Article history:
Received 9 June 2012
Received in revised form
10 November 2012
Accepted 31 December 2012
Available online 23 January 2014

Keywords:
Semi-supervised learning
Latent topic model
Object classification

ABSTRACT

Latent topic models are applied to analyze the low-dimensional semantic meaning of documents and images, which are widely used in object categorization. However, the unsupervised topic model cannot guarantee that the learned topics have a good relation with class labels, while manually aligning and labeling all training images are expensive and subjective in real applications. Aiming at using a small amount of partial labels to find topics much more suitable for classification, joint distribution from multi-conditional learning is adopted in this paper to generate semi-supervised topic models. Semi-supervised LDA and pLSA models are proposed when the joint distribution is known or partially known. Experimental results on natural scene categorization and head pose classification tasks show that the proposed method remains promising using only partial labels in the training process, which demonstrates the effectiveness of the proposed method.

1. Introduction

Conventionally, the classification methods can be grouped into generative methods and discriminative ones. Generative methods estimate the joint distribution of classes and the observed data. While the discriminative methods focus on the conditional probability of the class label given the observed data and model parameters, in which the classification usually is determined by the decision boundary. The classification accuracy of those two kinds of models is varied with the amount of training data. With adequate training data, the generative methods have higher classification accuracy than the discriminative ones. However, when the size of the training data is small, the discriminative model usually has better performance [1].

The generative topic models, such as probabilistic latent semantic analysis (pLSA) [2], latent Dirichlet allocation (LDA) [3] and their variations, aim to discover the latent semantic topics in a large collection of documents with statistical learning. With the bag of words (BoW) representation of images, these models are widely used in computer vision fields such as human action and image scene categorization [4–9]. LDA and pLSA models assume that each word is generated from the latent topic [3]. However, the semantic meaning of each topic is usually unclear since they are fully unsupervised. Thus, when applied to classification, the correspondence between topics and class labels is unknown. To link the class information with the latent topics, some supervised topic models were proposed.

Supervised topic models can be roughly divided into document-level and word-level supervised methods. In the first case, the class label is linked to the mixture of topics in the directed graphical model. Bayesian hierarchical model [4] is a classic method. While in the word-level supervised methods, the class label is connected to the topic variable, such as the labeled LDA model [10], supervised LDA(SLDA) [11]. Compared with document-level methods, word-level supervised methods are much more flexible. SLDA is a commonly used word-level supervised method, in which the topic variable is conditionally depending on words and class labels. In [12], Softmax function between topic and class label was used for classification task.

In the supervised topic models, such as SLDA, MedLDA [13], and DiscLDA [14] *etc.*, the training data should be labeled. Although more labels can improve the discriminative ability of the model, class labeling is tedious and subjective in real applications. It leads to a significant interest in semi-supervised topic models [15–18]. A straightforward way of adding partially labeled data in the generative model is to use the expectation maximization (EM) algorithm. In the training process, a pseudo label is assigned to the unlabeled document by initialization. Then the EM algorithm is applied until the optimum label is obtained. The performance of this method is sensitive to the initialization and expensive in calculation complexity. Reference [15] incorporates the partial domain knowledge to constrain the "must-link" and "cannot link" between words in the generative model. Steyvers et al. introduce the semantic concepts into the LDA model to train them jointly with the unlabeled words. However, these two methods focus on building relations among words. The relationship between the latent topic and the class label is still unclear [16]. A semi-latent Dirichlet allocation method is proposed in [17]. It assumes that each word is corresponding to an image and the topic of visual

* Corresponding author: Tel.: +86 029 88431533.
E-mail addresses: weiweinwpu@nwpu.edu.cn, weiwin1979@gmail.com (W. Wei).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.12.077

word is observable. Then the value of topic is set as class label in the learning process. Compared with the LDA model, this semi-latent topic model can achieve better classification result for video-based human action recognition. However, this method cannot be used if we cannot assign each word a class label, which is common in image classification. Using a regularization framework, Zhuang et al. propose a semi-supervised latent Dirichlet allocation model [18]. But this model is only applicable to binary classification, the multi-class classification method is not involved in this paper. Different from the above methods, we want to propose a new semi-supervised topic model benefiting from exist unsupervised and supervised topic models, which can be used for multi-class image classification.

One trend of semi-supervised classification is to hybrid the supervised and unsupervised methods by taking the advantage of both methods [1]. Multi-conditional learning (MCL) is one of the well-known hybrid methods [19]. In MCL, the hybrid discriminative and generative models are generated from their joint likelihood function, which share the same model parameter. We adopt the similar method in this paper to hybrid the unsupervised and supervised topic models, which are both generative models. According to the joint distribution is known or partially known, the hybrid semi-supervised LDA(HSLDA) model and the hybrid semi-supervised pLSA(HSpLSA) model are proposed in this paper. We aim to use a small amount of labels to help the generative topic model find semantic topics much more suitable for classification. We validate the proposed method on natural scene categorization and head pose classification tasks. Experimental results demonstrate the effectiveness of the proposed method.

2. Framework of hybrid semi-supervised topic model

To utilize the advantages of both the supervised and unsupervised models, we propose a jointly distributed semi-supervised topic generation method in this section. We denote a partially labeled corpus of documents as $D=\{D_L, D_U\}$, where $D_L=\{(\boldsymbol{w}_1, y_1), ..., (\boldsymbol{w}_l, y_l)\}$ is a set of labeled documents. $\boldsymbol{w}_d$ is the d-th labeled document. y_d is the class label of $\boldsymbol{w}_d$. Similarly, $D_U=\{\boldsymbol{w}_{l+1}, ..., \boldsymbol{w}_{l+u}\}$ represents the unlabeled documents. Similar to the MCL method [19], the joint distribution of the labeled and unlabeled documents can be represented as Eq.(1). In the hybrid semi-supervised topic model, MCL is used to hybrid two generative models.

$$o(\Theta)=p(D|\Theta)=p(D_L|\Theta)^{\lambda_1}p(D_U|\Theta)^{\lambda_2} \tag{1}$$

Due to the independence of each document, $p(D_L|\Theta)$ and $p(D_U|\Theta)$ can be represented by

$$\begin{aligned} p(D_L|\Theta)&=\prod_{d=1}^{l} p(\boldsymbol{w}_d, y_d|\Theta) \\ p(D_U|\Theta)&=\prod_{d=l+1}^{l+u} p(\boldsymbol{w}_d|\Theta) \end{aligned} \tag{2}$$

Θ is the model parameter, $p(D_L|\Theta)$ and $p(D_U|\Theta)$ are the likelihood functions of the supervised and unsupervised topic models. λ_1 and λ_2 are the weights of each model.

To simplify the calculation, the log likelihood of $o(\Theta)$ is obtained as

$$\log\ o(\Theta)=\lambda_1 \log\ p(D_L|\Theta)+\lambda_2 \log\ p(D_U|\Theta) \tag{3}$$

The model parameter Θ can be learned by

$$\widehat{\Theta}=\arg\max_{\Theta}(\log\ o(\Theta)) \tag{4}$$

In this paper, based on the above joint distribution, we propose two semi-supervised topic models. In case one, the distribution of both unsupervised and supervised topic models are given. How to generate the semi-supervised one using the joint distribution is proposed in Section 3. In case two, only the distribution of the unsupervised topic model is given. How to generate a semi-supervised topic model is demonstrated in Section 4.

3. The hybrid semi-supervised LDA model

LDA and SLDA models are unsupervised and supervised methods, respectively. In this section, based on these two known models, we use the joint distribution, as Eq. (3) shows to generate the HSLDA model. The variational inference method is adopted to accomplish the inference and part of the learning procedures for the HSLDA model. In this section, we first introduce LDA, SLDA and the variational inference method. Then the HSLDA model and its learning and inference methods are given.

3.1. Brief introduction of LDA, SLDA and the variational inference method

LDA model is a generative unsupervised topic model. It is proposed to model text corpora by discovering the latent semantic topics in a large collection of text documents. In the LDA model, topic is characterized by a particular distribution over vocabulary words and each document is characterized by a random mixture of topics indicating the probability of each topic [3].

LDA model can be represented in a graphical model as shown in Fig. 1(a). Suppose there are M documents, each has N words (we use N for notation simplicity although the words number for each document can be different), with K topics and a vocabulary of size V. In this model, α and β are corpus-level parameters. Parameter α is the Dirichlet prior on the per-document topic distributions and controls how the mixture of topic θ_i for each document is generated. β is a parameter of the multinomial distribution, which encodes each of the K topics as a distribution over vocabulary V. The words, w_{ij}, $i\in(1,2,...,M)$, $j\in(1,2,...,N)$ in the corpus are observations. The mixture topic weights θ_i for each document and the probability of topic z_{ij} for each word are the variables need to be inferred in this model.

Each observed document is composed of rich words. Considering the words are indexed from a vocabulary, each document can be represented as a high dimensional word-occurrence vector. However, document classification in the high dimensional space is usually difficult and inaccurate. Thus, given the observation $\mathbf{w}$ and the corpus-level parameters α and β, we want to infer $p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$ in the LDA model [3]. We name it as the inference task, which can be used for document classification in the latent low-dimensional topic space. While the learning task is to determine the optimum parameters of α and β that maximize the likelihood of all observations [3].

In LDA, $p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$ is represented as

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)=\frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \tag{5}$$

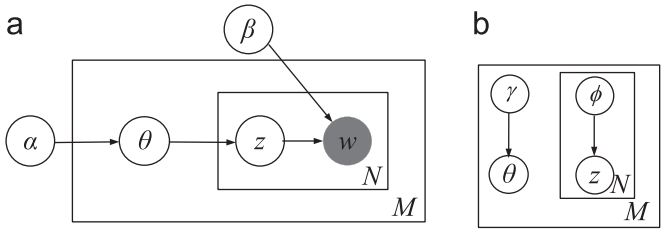

Fig. 1. The demonstration of directed graphical models of LDA. (a) The graphical model of LDA, (b) The simplified variational graphical model to approximate the LDA model.