Engineering Applications of Artificial Intelligence 35 (2014) 176–186

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

A variational Bayes model for count data learning and classification

Ali Shojaee Bakhtiari [a], Nizar Bouguila [b,*]

[a] Department of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada H3G 1T7
[b] The Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, QC, Canada H3G 1T7

ARTICLE INFO

Article history:
Received 21 September 2013
Received in revised form
27 March 2014
Accepted 29 June 2014
Available online 20 July 2014

Keywords:
Latent topic models
Count data
Generalized Dirichlet distribution
Variational Bayes
Text classification
Visual scene categorization

ABSTRACT

Several machine learning and knowledge discovery approaches have been proposed for count data modeling and classification. In particular, latent Dirichlet allocation (LDA) (Blei et al., 2003a) has received a lot of attention and has been shown to be extremely useful in several applications. Although the LDA is generally accepted to be one of the most powerful generative models, it is based on the Dirichlet assumption which has some drawbacks as we shall see in this paper. Thus, our goal is to enhance the LDA by considering the generalized Dirichlet distribution as a prior. The resulting generative model is named latent generalized Dirichlet allocation (LGDA) to maintain consistency with the original model. The LGDA is learned using variational Bayes which provides computationally tractable posterior distributions over the model's hidden variables and its parameters. To evaluate the practicality and merits of our approach, we consider two challenging applications namely text classification and visual scene categorization.

1. Introduction

Count data appear in many domains (e.g. data mining, computer vision, machine learning, pattern recognition, and bioinformatics) and applications. Examples include textual documents and images modeling and classification where each document or image can be represented by a vector of frequencies of words (Nigam et al., 2000) or visual words (Csurka et al., 2004), respectively. The extraction of knowledge hidden in count data is a crucial problem which has been the topic of a significant amount of research in the past. The naive Bayes assumption, through the consideration of the multinomial distribution, was extensively used for count data modeling (Nigam et al., 2000). However, serious deficiencies such as, being prone to training bias, the need for the assumption of independence for features and failure to model text well, were observed with the application of the multinomial distribution as thoroughly discussed in Madsen et al. (2005) and Bouguila and Ziou (2007a). The most widely used solution to overcome these deficiencies is the consideration of the Dirichlet distribution as a conjugate prior to the multinomial which generally offers better flexibility, generalization and modeling capabilities (Madsen et al., 2005; Bouguila and Ziou, 2007a; Mei et al., 2007). Despite many favorable features, it has been pointed out that the Dirichlet distribution has some shortcomings, also. The main disadvantages of the Dirichlet distribution are its very restrictive negative covariance matrix and the fact that the elements with similar mean values must have absolutely the same variance which is not always the case in real-life applications (Bouguila, 2008). To overcome those deficiencies, research has been focused on providing a transition from the Dirichlet assumption to better modeling assumptions (Bouguila, 2011). The context of this paper is majorly about this transition as well, where the ultimate goal is to have more accurate data modeling.

One of the immediate applications of proper data modeling is classification. It covers a vast extent of problems such as placement of textual data into appropriate library entries or classifying objects into their relevant categories. In this context, one of the most challenging tasks is the classification of visual scenes without going deep inside their semantics. The challenge behind the former is that visual scenes are generally composed of a huge number of minute objects. The presence of this ever occurring objects makes it extremely complicated to develop useful classifiers based on the semantics alone. After all one would expect to see roads, trees, sun and the sky recurring in scenes both taken inside the city or in the suburb. The need to consider the presence of recurring data singletons, whether words, visual words or visual objects, led to the so-called topic based models. Latent semantic indexing (LSI) (Deerwester et al., 1990) is the first successful model proposed to extract recurring topics from data. It was proposed for textual documents modeling using mainly singular value decomposition (SVD). A generative successful extension of LSI called probabilistic latent semantic indexing (PLSI) was proposed in Hofmann (2001). And a hierarchical extension of PLSI was proposed in Vinokourov and

* Corresponding author. Tel.: +1 5148482424; fax: +1 5148483171.
E-mail addresses: al_sho@encs.concordia.ca (A.S. Bakhtiari), nizar.bouguila@concordia.ca (N. Bouguila).

http://dx.doi.org/10.1016/j.engappai.2014.06.023
0952-1976/

Girolami (2002). However, PLSI is only generative at the words layer and does not provide a probabilistic model at the level of documents. Therefore, two major problems arise with PLSI. Firstly, the number of parameters increases with the number of documents. Secondly, it is not clear how one can learn a document outside of the training phase. To overcome these shortcomings, the authors in Blei et al. (2003a) proposed the LDA model which has so far proven to be a reliable and versatile approach for data modeling. LDA has received a particular attention in the literature and several applications (e.g. natural scene classification Fei-Fei and Perona, 2005) and extensions have been proposed. Examples of extensions include the hierarchical version of LDA (Blei et al., 2003b), used for instance in Sivic et al. (2008) for hierarchical object classification, the online version proposed in Hoffman et al. (2010), and the discriminative supervised version described in Lacoste-Julien et al. (2008). Of course, these extension efforts are useful for several real-life applications and scenarios, but have ignored an important aspect of LDA namely the fact that it considers the Dirichlet distribution, including its drawbacks, for generating latent topics. Previously other researchers tried to develop latent topic models based on the conjugate priors other than Dirichlet (Caballero et al., 2012). Their model however is based on Gibbs sampling and Markov chain Monte Carlo (MCMC) method (Robert and Casella, 2004). The advantage of the MCMC method is its relative ease of derivation. However, it has been shown that sampling methods require much more computation time than deterministic methods such as variational Bayes. Therefore, where it is possible to derive an analytic form, deterministic models are more preferable. In this work we shall focus on deriving an extension to the LDA model using the generalized Dirichlet assumption using the variational Bayes method.

Recently, the second author has shown that the generalized Dirichlet is a good alternative to the Dirichlet when using finite mixture models for count data clustering (Bouguila, 2008). Like the Dirichlet, the generalized Dirichlet distribution is a conjugate prior to the multinomial distribution which is a crucial property in the LDA model. Moreover, the generalized Dirichlet has a more versatile covariance matrix and also it lifts the variance limitations facing Dirichlet vectors (Bouguila, 2008). The goal of this work is to propose an extension of LDA based on the generalized Dirichlet distribution. To maintain consistency with the LDA model we call our model, latent generalized Dirichlet allocation (LGDA). We shall develop a variational Bayes estimation approach inspired from the one proposed in Blei et al. (2003a), yet with the generalized Dirichlet assumption. The Dirichlet distribution is a special case of the generalized Dirichlet distribution (Connor and Mosimann, 1969; Bouguila and Ziou, 2007b), therefore it is expectable that the LGDA will provide good modeling capabilities. In the experimental results we shall elaborate the conjunctions between the two models further. We shall compare the two models via two challenging applications namely text and visual scene classification.

The rest of the paper is organized as follows. In Section 2, we introduce the LGDA model and give the detailed derivations to learn its parameters. Section 3 is devoted to the presentation of the results of applying both LDA and LGDA. The applications concern text and visual scene classification and are used to show the strengths and weaknesses of both models. Finally, conclusion and some thoughts about future directions follow in Section 4.

2. Latent generalized Dirichlet allocation

2.1. The model

Like LDA, LGDA is a fully generative probabilistic model over a corpus. A corpus in our case is a collection of M documents (or images) denoted by $\boldsymbol{M}=(\boldsymbol{w_1},\boldsymbol{w_2},\ldots,\boldsymbol{w_M})$. And each document $\boldsymbol{w}_m$ is a sequence of N_m words $\boldsymbol{w}_m=(w_{m1},\ldots,w_{mN_m})$. In what follows, for sheer convenience, we drop the index m wherever we are not referring to a specific document. The word $w_n=(w_n^1,\ldots,w_n^V)$ is considered as a binary vector drawn from a vocabulary of V words, so that $w_n^j=1$ if the j-th word is chosen and zero, otherwise. The model proceeds with generating every single word (or visual word) of the document (or the image) through the following steps:

1. Choose $N\propto Poisson(\zeta)$.
2. Choose $(\theta_1,\ldots,\theta_d)\propto GenDir(\vec{\xi})$.
3. For each of the N words w_n:
   (a) choose a topic $z_n\propto Multinomial(\vec{\theta})$,
   (b) choose a word w_n from $p(w_n|z_n,\mu_w)$.

In above z_n is a $d+1$ dimensional binary vector of topics defined so that $z_n^i=1$ if the i-th topic is chosen and zero, otherwise. We define $\vec{\theta}=(\theta_1,\ldots,\theta_{d+1})$, where $\theta_{d+1}=1-\sum_{i=1}^{d}\theta_i$. We define matrix μ_w so that a chosen topic is attributed to a multinomial μ_w over the vocabulary of words so that $\mu_{w(ij)}=p(w^j=1|z^i=1)$, from which every word is randomly drawn. $p(w_n|z_n,\mu_w)$ is a single draw multinomial probability conditioned on z_n and $GenDir(\vec{\xi})$ is a d-variate generalized Dirichlet distribution with parameters $\vec{\xi}=(\alpha_1,\beta_1,\ldots,\alpha_d,\beta_d)$ and probability distribution function given by

$$p(\theta_1,\ldots,\theta_d|\vec{\xi})=\prod_{i=1}^{d}\frac{\Gamma(\alpha_i+\beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)}\theta_i^{\alpha_i-1}\left(1-\sum_{j=1}^{i}\theta_j\right)^{\gamma_i} \tag{1}$$

where $\gamma_i=\beta_i-\alpha_{i+1}-\beta_{i+1}$. It is straightforward to show that when $\beta_i=\alpha_{(i+1)}+\beta_{(i+1)}$, the generalized Dirichlet distribution is reduced to Dirichlet distribution (Bouguila and Ziou, 2007b). We define $\vec{\theta}=(\theta_1,\ldots,\theta_{d+1})$, where $\theta_{d+1}=1-\sum_{i=1}^{d}\theta_i$. With the above parameters, the mean and the variance matrix of the generalized Dirichlet elements are as follows (Bouguila and Ziou, 2007b):

$$E(\theta_i)=\frac{\alpha_i}{\alpha_i+\beta_i}\prod_{k=1}^{i-1}\frac{\beta_k}{\alpha_k+\beta_k} \tag{2}$$

$$Var(\theta_i)=E(\theta_i)\left(\frac{\alpha_i+1}{\alpha_i+\beta_i+1}\prod_{k=1}^{i-1}\frac{\beta_k+1}{\alpha_k+\beta_k}+1-E(\theta_i)\right) \tag{3}$$

and the covariance between θ_i and θ_j is given by

$$Cov(\theta_i,\theta_j)=E(\theta_j)\left(\frac{\alpha_i}{\alpha_i+\beta_i+1}\prod_{k=1}^{i-1}\frac{\beta_k+1}{\alpha_k+\beta_k}+1-E(\theta_i)\right) \tag{4}$$

It can be seen from Eq. (4) that the covariance matrix of the generalized Dirichlet distribution is more general than the covariance matrix of the Dirichlet distribution and unlike Dirichlet distribution it is possible for two elements inside the random vector to be positively correlated. Also unlike Dirichlet two elements with the same mean value can have different variances. Generalized Dirichlet distribution, like the Dirichlet distribution, belongs to the exponential family of distributions (see Appendix A). This means that the generalized Dirichlet distribution has a conjugate prior that can be developed in a formal way, which is an important property that we shall used in the following for the learning of our model. It turns out also that generalized Dirichlet like Dirichlet is the conjugate prior of the multinomial distribution. This implies that if $(\theta_1,\ldots,\theta_d)$ follows a generalized Dirichlet distribution with parameters $\vec{\xi}$, and $\vec{N}=(n_1,\ldots,n_{d+1})$ follows a multinomial with parameter $\vec{\theta}$, then the posterior distribution $p(\vec{\theta}\,|\,\vec{\xi},\vec{N})$ also follows a generalized Dirichlet distribution with parameters ξ' given as follows (Bouguila, 2008):

$$\alpha_i'=\alpha_i+n_i \tag{5}$$

$$\beta_i'=\beta_i+\sum_{l=i+1}^{d+1}n_l \tag{6}$$

Having our generalized Dirichlet prior in hand, we proceed with defining the $(d+1)\times V$ word-topic probability matrix μ_w which element $\mu_{w_{ij}} = p(w_j = 1|z_i = 1)$ shows the probability of drawing the j-th word given that the i-th latent topic is chosen. Like the LDA case, we proceed with assuming a non-generated μ_w matrix, but we will show that this assumption does not have a serious impact and it can be revoked without bringing harm to the entire model. By assuming conditional independence of the variables, the same as LDA, one can deduce the following joint distribution:

$$p(\vec{\theta}, \boldsymbol{z}, \boldsymbol{w}|\vec{\xi}, \mu_w) = p(\vec{\theta}|\vec{\xi})p(\boldsymbol{w}|\boldsymbol{z}, \mu_w)p(\boldsymbol{z}|\vec{\theta}) \tag{7}$$

where $\boldsymbol{z}$ is the set of latent topics. Integrating over the $\vec{\theta}$ parameters and the topic space gives

$$p(\boldsymbol{w}|\vec{\xi}, \mu_w) = \int \prod_{i=1}^{d} \frac{\Gamma(\alpha_i+\beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} \theta_i^{\alpha_i-1} \left(1 - \sum_{j=1}^{i} \theta_j\right)^{\gamma_i} \prod_{n=1}^{N} \sum_{i=1}^{d+1} \prod_{j=1}^{V} (\theta_i \mu_{w_{ij}})^{w_n^j} \, d\vec{\theta} \tag{8}$$

In the previous equation, $\vec{\xi}$ and μ_w are the corpus level parameters that are selected once per each document in the corpus. $\vec{\theta}$ is the document level parameter and is chosen once per document. $\boldsymbol{z}$ and $\boldsymbol{w}$ are word level parameters and are chosen once per every word inside each document. Thus, we can obtain the probability of the corpus as follows:

$$p(\boldsymbol{D}|\vec{\xi}, \mu_w) = \prod_{m=1}^{M} p(\boldsymbol{w}_m|\vec{\xi}, \mu_w) \tag{9}$$

LGDA has basically the same probabilistic graphical model as LDA as it is shown in Fig. 1.

2.2. LGDA inference

The main inference problem of LGDA is estimating the posterior of the hidden variables, $\vec{\theta}$ and $\boldsymbol{z}$:

$$p(\vec{\theta}, \boldsymbol{z}|\boldsymbol{w}, \vec{\xi}, \mu_w) = \frac{p(\vec{\theta}, \boldsymbol{z}, \boldsymbol{w}|\vec{\xi}, \mu_w)}{p(\boldsymbol{w}|\vec{\xi}, \mu_w)} \tag{10}$$

The above equation is known to be intractable. As proposed in Blei et al. (2003a), an efficient way to estimate the parameters in this intractable posterior is to use the variational Bayes (VB) inference. VB inference offers a solution to the intractability problem by determining a lower bound on the log likelihood of the observed data which is mainly based on considering a set of variational distributions on the hidden variables (Jordan et al., 1999; Watanabe and Watanabe, 2006; Fan et al., 2012):

$$q(\vec{\theta}, \boldsymbol{z}|\boldsymbol{w}, \vec{\xi}_q, \Phi_{\boldsymbol{w}}) = q(\vec{\theta}|\vec{\xi}_q) \prod_{n=1}^{N} q(z_n|\phi_n) \tag{11}$$

In the above $q(\vec{\theta}|\vec{\xi}_q)$ can be viewed as a variational generalized Dirichlet distribution, calculated once per document, $q(z_n|\phi_n)$ is a multinomial distribution with parameter ϕ_n extracted once for every single word inside the document, and $\Phi_{\boldsymbol{w}} = \{\phi_1, \phi_2, \ldots, \phi_N\}$. Using Jensen's inequality (Jordan et al., 1999) one can derive the following:

$$\log p(\boldsymbol{w}|\vec{\xi}, \mu_w) \geq E_q[\log p(\vec{\theta}, \boldsymbol{z}, \boldsymbol{w}|\vec{\xi}, \mu_w)] - E_q[\log q(\vec{\theta}, \boldsymbol{z})] \tag{12}$$

Assigning $L(\vec{\xi}_q, \Phi_{\boldsymbol{w}}; \vec{\xi}, \mu_w)$ to the right-hand side of the above equation it can be shown that the difference between the left-hand side and the right-hand side of the equation is the KL divergence between the variational posterior probability and the actual posterior probability, thus we have

$$\log p(\boldsymbol{w}|\vec{\xi}, \mu_w) = L(\vec{\xi}_q, \Phi_{\boldsymbol{w}}; \vec{\xi}, \mu_w) + KL(q(\vec{\theta}, \boldsymbol{z}|\vec{\xi}_q, \Phi_{\boldsymbol{w}}) || p(\vec{\theta}, \boldsymbol{z})|\boldsymbol{w}, \vec{\xi}, \mu_w)) \tag{13}$$

The left-hand side of the above equation is constant in relation to variational parameters, therefore to minimize the KL divergence on the right-hand side one can proceed with maximizing $L(\vec{\xi}_q, \Phi_{\boldsymbol{w}}; \vec{\xi}, \mu_w)$. Up to here the formulation basically follows the LDA model. The divergence of the models begins when we proceed with assigning the generalized Dirichlet distribution as the parameter generator instead of the LDA Dirichlet assumption. In Appendix B we bring the breakdown of $L(\vec{\xi}_q, \Phi_{\boldsymbol{w}}; \vec{\xi}, \mu_w)$.

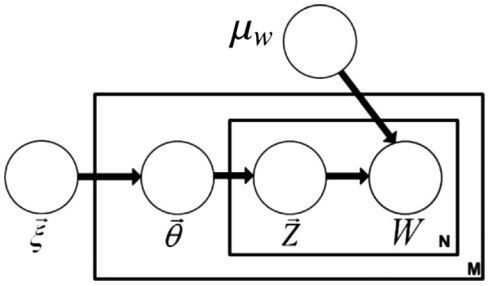

Fig. 1. Graphical representation of LGDA model. The shaded circles show observed nodes. The blank circles are the hidden nodes. From outside to inside is the corpus space, the document space and the word space.

Using variational inference to maximize the lower bound $L(\vec{\xi}_q, \Phi_{\boldsymbol{w}}; \vec{\xi}, \mu_w)$ with respect to ϕ_{nl}, we derive the following updating equations for the variational multinomial (see Appendix B.1)

$$\phi_{nl} = \beta_{lv} e^{(\lambda_n - 1)} e^{(\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l))} \tag{14}$$

$$\phi_{n(d+1)} = \beta_{(d+1)v} e^{(\lambda_n - 1)} e^{(\Psi(\delta_d) - \Psi(\gamma_d + \delta_d))} \tag{15}$$

where Ψ is the digamma function, $\beta_{lv} = p(w^v = 1|z^l = 1)$ and the weighing constant $e^{\lambda_n - 1}$ is given by

$$e^{\lambda_n - 1} = \frac{1}{\sum_{l=1}^{d} \beta_{lv} e^{(\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l))} + \beta_{(d+1)v} e^{(\Psi(\delta_d) - \Psi(\gamma_d + \delta_d))}} \tag{16}$$

Maximizing the lower bound L with respect to the variational generalized Dirichlet parameter gives the following updating equations (see Appendix B.1):

$$\gamma_l = \alpha_l + \sum_{n=1}^{N} \phi_{nl} \tag{17}$$

$$\delta_l = \beta_l + \sum_{n=1}^{N} \sum_{ll=l+1}^{d+1} \phi_{n(ll)} \tag{18}$$

Comparing the above equations with Eqs. (5) and (6) shows that the variational generalized Dirichlet for each document acts as a posterior in the presence of the variational multinomial parameters. The same conclusion was observed in Blei et al. (2003a) for the LDA case. This is a direct result of the conjugacy between the generalized Dirichlet and the multinomial distribution.

2.3. Parameters estimation

The goal of this subsection is to find the model's parameters estimates based on the variational parameters derived in the last subsection. One needs to consider that the LGDA parameters are corpus parameters and therefore they are estimated by considering all M documents inside the corpus. In the following, we denote $L = \sum_{m=1}^{M} L_m$ as the lower bound corresponding to all the corpus, where L_m is the lower bound corresponding to each document m.

Maximizing the corpus lower bound L with respect to $\mu_{w(lj)}$ delivers the following updating equation (see Appendix B.3):

$$\mu_{w(lj)} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dnl} w_{dn}^j \tag{19}$$

The model's parameters are the last ones to be derived. Following the work of Minka (2007), it was shown in Blei et al. (2003a) that in order to derive LDA parameters it was feasible to use the Newton–Raphson algorithm for parameters estimation. It was also shown that due to the characteristics of the Dirichlet distribution, it is possible to exchange the computationally demanding problem of inverting the Hessian matrix of the Lower bound with a linear operator and therefore reducing the model complexity.

The Hessian matrix of the generalized Dirichlet distribution offers the same useful, albeit in a different way, simplification. This characteristic was analyzed in Bouguila and Ziou (2007b). The nature of the generalized Dirichlet distribution leads the Hessian matrix to take a 2×2 block-diagonal shape. The inverse matrix of a block-diagonal matrix is another block-diagonal matrix consisting of the inverses of the blocks of the original matrix. Therefore the problem of inverting the $2d \times 2d$ Hessian matrix is reduced to computing the inverse of 2×2 matrix for d instances. The complete derivation of the model parameters is brought in Appendix B.4.

The last formulation that we need to derive to prepare our model for the classification task is the likelihood of a document in our model. This can be done by deriving first the likelihood of a randomly chosen word w_n inside the document:

$$\begin{aligned} p(w_n|\vec{\xi}) &= \sum_{l=1}^{d+1} \int p(w_n|z_l)p(z_l|\vec{\theta})p(\vec{\theta}|\vec{\xi})\, d\vec{\theta} \\ &= \sum_{l=1}^{d+1} p(w_n|z_l) \int p(z_l|\vec{\theta})p(\vec{\theta}|\vec{\xi})\, d\vec{\theta} \\ &= \sum_{l=1}^{d} \beta_{l(v|w_n^v=1)} E[\theta_l] + \beta_{(d+1)(v|w_n^v=1)} \left(1 - \sum_{l=1}^{d} E[\theta_l]\right) \end{aligned} \tag{20}$$

Combining Eq. (2) with Eq. (20) delivers the formulation for the word likelihood as follows:

$$\begin{aligned} p(w_n|\vec{\xi}) = &\sum_{l=1}^{d} \beta_{l(v|w_n^v=1)} \left(\frac{\alpha_l}{\alpha_l+\beta_l} \prod_{k=1}^{l-1} \frac{\beta_k}{\alpha_k+\beta_k}\right) \\ &+ \beta_{(d+1)w_n} \left(1 - \sum_{l=1}^{d} \left(\frac{\alpha_l}{\alpha_l+\beta_l} \prod_{k=1}^{l-1} \frac{\beta_k}{\alpha_k+\beta_k}\right)\right) \end{aligned} \tag{21}$$

The log likelihood of a document $\boldsymbol{w}_m$ is derived as the sum of the log likelihoods of the words present inside the document and

Table 1
Extracted classes and number of available documents per each class.

Class name	Number of documents
'acq'	2293
'crude'	579
'earn'	3939
'grain'	593
'interest'	479
'money-fx'	729

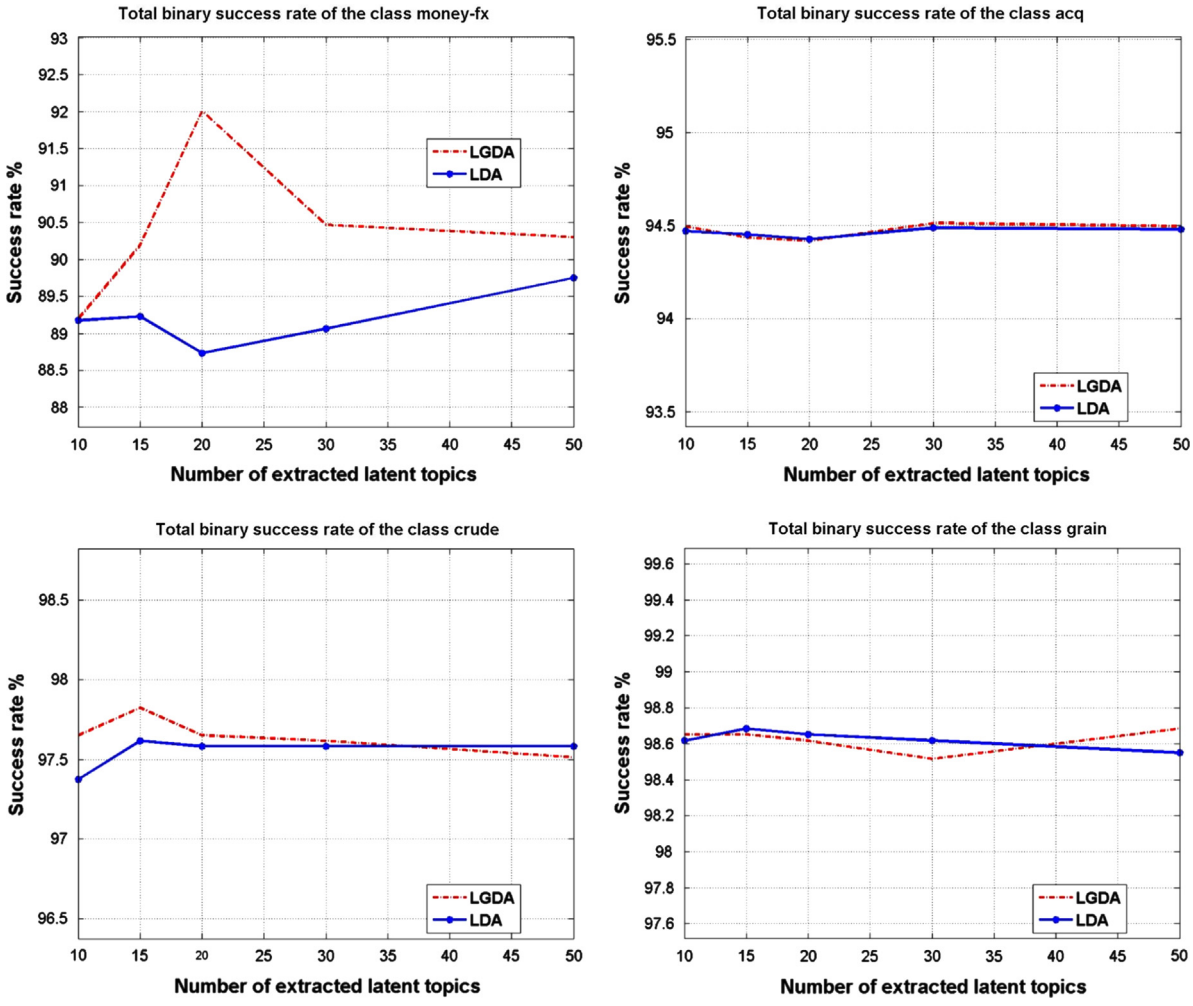

Fig. 2. Examples of binary classification success rates of the LGDA and LDA models when applied for text classification. Red line: LGDA, blue line: LDA. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

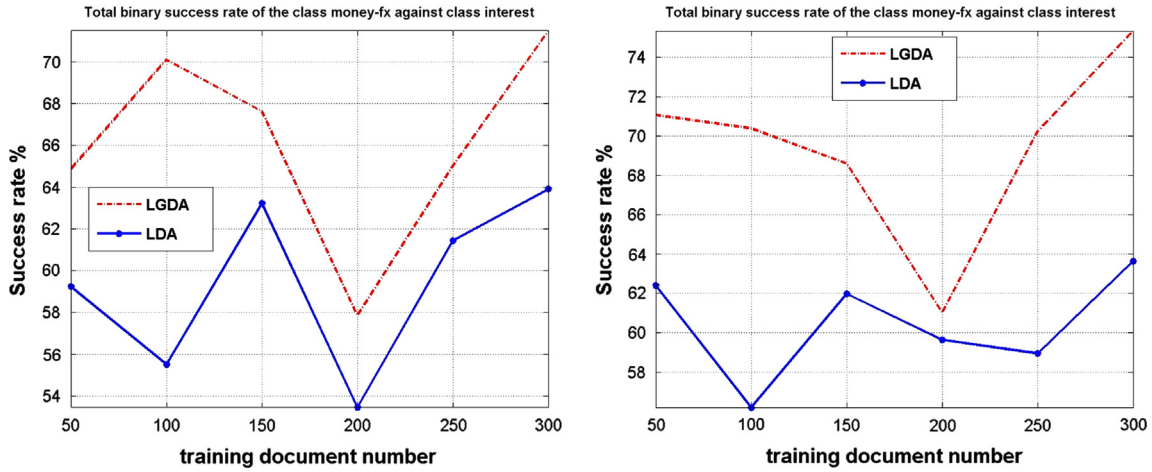

Fig. 3. Comparison of binary classification success rates of the LGDA and LDA models for 'Money-fx' class against 'interest' class when we consider (a) 15 extracted latent topics, and (b) 30 extracted latent topics. Red line: LGDA, blue line: LDA. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

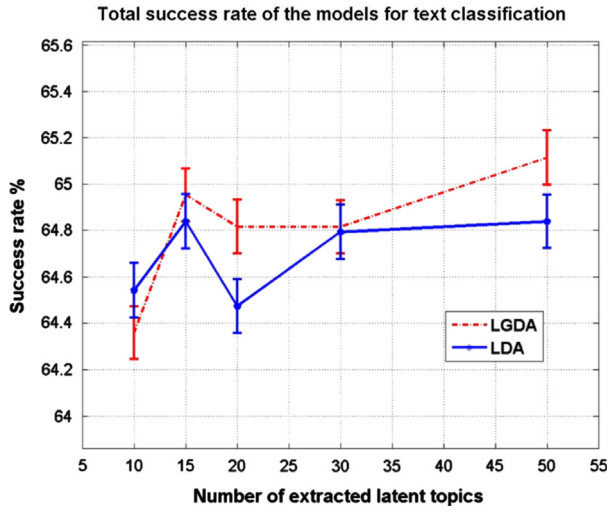

Fig. 4. Total text classification success rates obtained using LGDA and LDA models. Red line: LGDA, blue line: LDA. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Table 2
Confusion matrix of the LGDA model, in the optimal case, when applied to text classification.

	acq	crude	earn	grain	interest	money-fx
acq	718	16	171	0	31	8
crude	149	507	130	12	40	61
earn	25	17	445	6	8	20
grain	33	14	81	554	28	25
interest	26	11	80	6	252	238
money-fx	49	14	93	15	120	374

therefore we have

$$\log\ p(\boldsymbol{w}_m|\vec{\xi}) = \sum_{n=1}^{N_m} \log\ p(w_n|\vec{\xi}) = \sum_{v=1}^{V} cnt_{mv} \log\ p(w_v|\vec{\xi}) \qquad (22)$$

where for each document $\boldsymbol{w}_m$, cnt_{mv} is the number of times the v-th word is drawn. For the classification purposes, the class that gives the highest likelihood is chosen as the class the document belongs.

Table 3
Confusion matrix of the LDA model, in the optimal case, when applied to text classification.

	acq	crude	earn	grain	interest	money-fx
acq	720	16	170	2	31	11
crude	150	510	132	12	36	61
earn	26	14	446	6	8	23
grain	30	12	79	554	28	29
interest	26	11	80	5	273	267
money-fx	48	16	93	14	103	335

3. Experimental results

In this section, we bring the results of applying the LGDA model on two distinct challenging applications namely text and visual scene classification. The main goal of both applications is to compare the LGDA and LDA performances.

3.1. Text classification

In text classification, the problem at hand is deciding which distinctive class to assign a given document to Sebastiani (2002) and Bouguila and Ziou (2009). An effective classification can be used for different purposes such as retrieval, recommendation, filtering, and topic detection (Stokes and Carthy, 2001). This problem has been the topic of extensive research in the past (see, for instance, Sebastiani, 2002; Ruiz and Srinivasan, 2002; Zhang et al., 2013, and references therein) and can be looked upon from two distinct but related ways. Assuming that the number of classes is known, from a first perspective text classification can be viewed a binary categorization problem where the main task is to decide to which class we should assign the text given two distinctively chosen classes. The other way to look upon the problem is to decide how accurately can the model assign the proper class to a document at the presence of all other classes. We proceed with giving results for each of the two mentioned scenarios in the following.

For our simulations, we chose the Reuters-21578 dataset[1] (Joachims, 1998). This dataset consists of 21 578 documents and in total there are more than 20 000 words present inside it. Independent works have, either manually or automatically,

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/.

already classified most of the 21 578 documents into superseding categories (Joachims, 1998; Vinokourov and Girolami, 2002). Even though there are many extracted categories thus obtained, not all of them contain enough documents to be suitable for training and testing purposes. Thus, we limit ourselves to the top 6 categories extracted from the dataset. They, in total, comprise more than 9000 documents of the original dataset and nearly all the words present in the unabridged dataset. Table 1 describes the considered classes.

To examine the classification accuracy of the models, in the first step we choose a certain number of the documents in each of the classes as training documents. Next, we learn our models for each of the chosen training sets, for different numbers of latent topics to observe the effect of choosing them on the classification accuracy. Classification in this first experiment is regarded as a binary process meaning that a given document is presented to two different trained classes and the class that gives the higher likelihood is chosen as the document class. Selected success rates of the two models are brought in Fig. 2. From each class 100 documents are randomly chosen for training and we limited our test to the classes which at least contained 500 documents. An interesting observation regarding the two models can be deduced from this figure. The Reuters dataset consists of relatively short documents that are presented as extremely sparse count vectors over the entire vocabulary set. Both LDA and LGDA use variational Bayes inference as their core learning method, however the sparsity of the count vectors causes both models to basically provide the same fit over the training set. The result is that facing sparse vectors, the two models roughly offer the same success rate. This can be seen in Fig. 2. There is an exception to this observation. When the two classes are similar to each other, one may expect that the models fail to separate them as precisely as when the classes are dissimilar. In this case the model that offers the better fitting to its training set could offer better classification. An instance of related classes is 'interest' and 'money-fx'; the classification success rates obtained when using LGDA and LDA, in this case, are displayed in Fig. 3. This example shows that when there are similarities between distinct classes, LGDA offers a more accurate classification than LDA. Thus, we can conclude, according to Figs. 2 and 3, that in the majority of cases LGDA offers either comparable or improved results as compared to LDA. That again coincides with our expectation that LGDA acts like or better than LDA.

In Fig. 4, we compare the total classification accuracies of the two models. We need to emphasize the difference between Figs. 2 and 4. While Fig. 2 shows the class by class comparison of success rates, Fig. 4 shows the total success rate of the two models. In order to suppress the effects of over compensation by different classes in derivation of Fig. 4 we limited the number of documents in each class to 1000. Tables 2 and 3 show the confusion matrices of the LGDA and LDA models, respectively, for the optimal cases (i.e. corresponding to the maximum rates in Fig. 4).

3.2. Visual scenes classification

3.2.1. Methodology

In this set of experiments, we apply our LGDA model to the challenging task of visual scenes classification which is a crucial step in several applications such as image annotation and retrieval (Carneiro et al., 2007; Rashedi et al., 2013), and content-based images recommendation (Boutemedjet et al., 2007). The main goal is to compare the LGDA to the LDA which was considered for the same task in Fei-Fei and Perona (2005). It is noteworthy that some adaptations to the original LDA were proposed in Fei-Fei and Perona (2005), and the reader is then referred to this paper for more details, to make it applicable to scenes classification. The very same adaptations were included in the LGDA without a need for further assumptions. Indeed, the main idea that we use here is based on the description of scenes using visual words (Csurka et al., 2004). This approach has emerged over the past few years and received strong interest that is mainly motivated by the fact that many of the techniques previously proposed for text classification can be adopted for images categorization (Csurka et al., 2004; van Gemert et al., 2010). For the construction of the visual words vocabulary, we need first to extract local descriptors from a set of training images. In our case, we use the scale invariant feature transform (SIFT) descriptors (Lowe, 2004). The extracted features are then quantized through clustering (the K-Means algorithm in our case) and the obtained d clusters centroids are considered as our visual words. Having the visual vocabulary in hand, each image can be represented as a d-dimensional vector containing the frequency of each visual word in that image. In our experiment we take 7 classes from the natural scenes dataset introduced in Oliva and Torralba (2001) and we combine it with one indoor scenes class from Fei-Fei and Perona (2005). The 7 classes chosen from the data set described in Oliva and Torralba (2001) are coast, forest, highway, inside of cities, open country, street, and tall building, which contain respectively 361, 329, 261, 309, 411, 293, and 356 images, respectively. The class chosen from the data set proposed in Fei-Fei and Perona (2005) is the bedroom category which contains 217 images. Examples of images from the different considered classes are shown in Fig. 5.

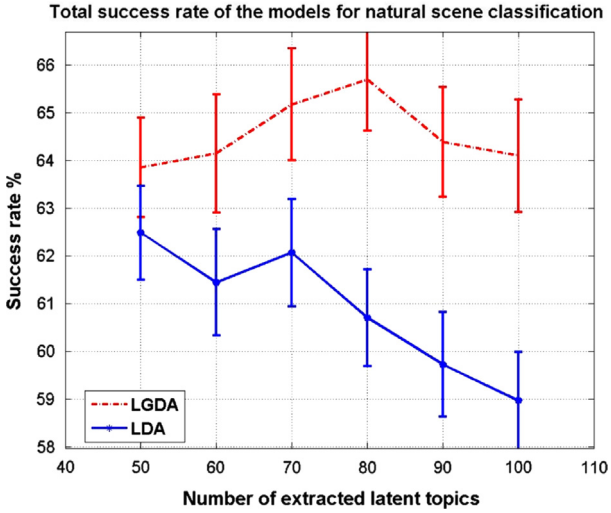

Fig. 6. Classification success rates, as a function of the number of extracted latent topics, of the LGDA and LDA models applied for the visual scenes classification task. Red line: LGDA, blue Line: LDA. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Fig. 5. Sample images from each group. (a) Highway, (b) Inside of cities, (c) Tall building, (d) Streets, (e) Forest, (f) Coast, (g) Open country, (h) Bedroom.

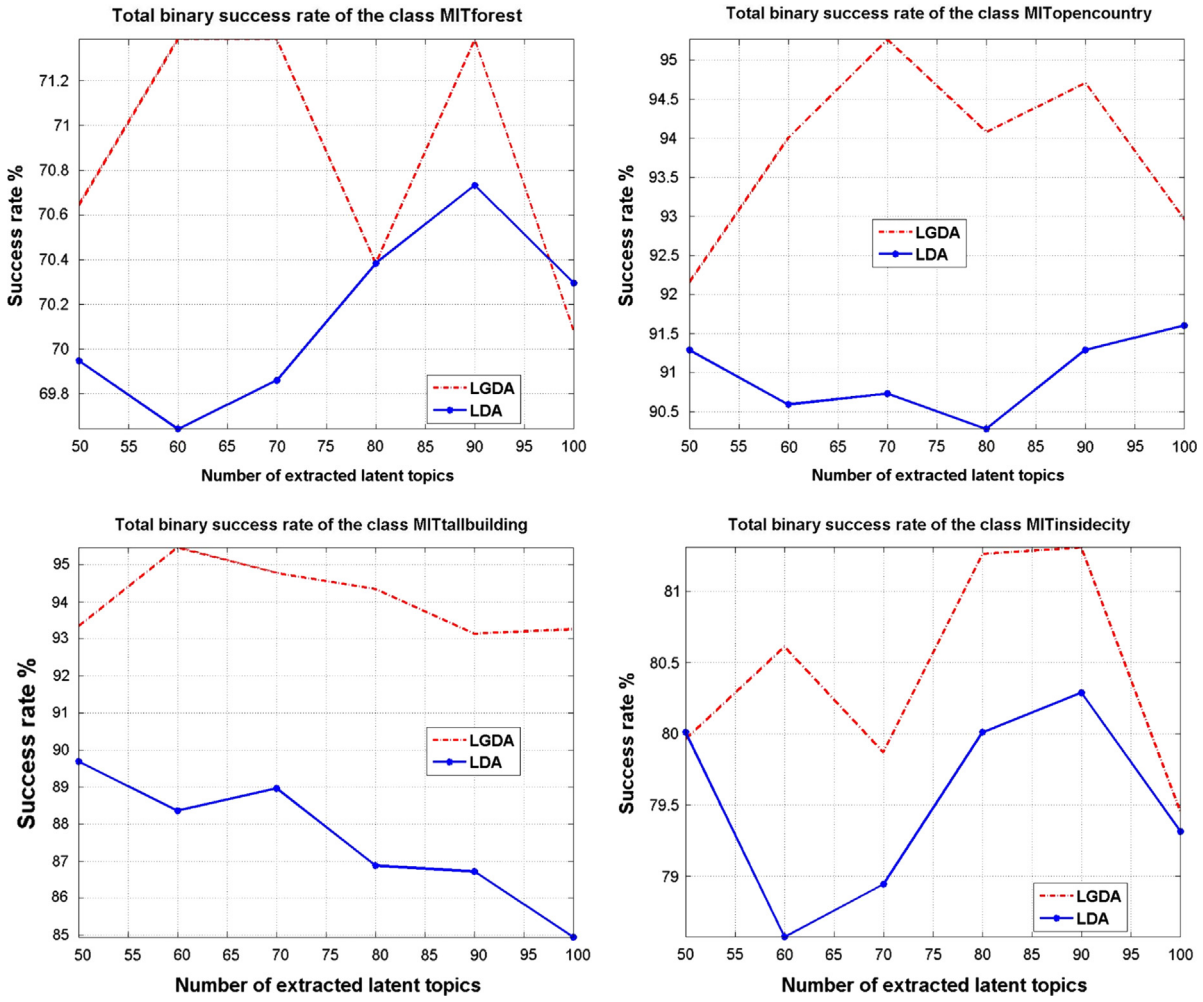

Fig. 7. Examples of per class classification success rates, as a function of the number of extracted latent topics, of the LGDA and LDA models. Red line: LGDA, blue Line: LDA. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Table 4
Optimal confusion matrix of the LGDA model applied for the scenes classification task.

	Coast	Forest	Highway	Inside of cities	Open country	Streets	Tall building	Bedroom
Coast	223	0	51	3	50	1	2	1
Forest	1	192	10	12	0	23	0	0
Highway	33	2	104	6	18	8	14	1
Inside of cities	0	25	8	173	0	26	0	0
Open country	76	30	7	13	321	10	18	11
Streets	0	57	41	15	0	173	0	0
Tall building	9	0	3	22	7	3	268	12
Bedroom	16	18	30	59	11	42	51	185

Table 5
Optimal confusion matrix of the LDA model applied for the scenes classification task.

	Coast	Forest	Highway	Inside of cities	Open country	Streets	Tall building	Bedroom
Coast	231	0	59	2	53	1	1	1
Forest	1	187	5	16	0	21	0	0
Highway	55	10	100	21	47	19	62	4
Inside of cities	0	21	6	168	0	22	0	0
Open country	49	27	5	7	289	6	13	7
Streets	1	53	49	16	0	178	0	0
Tall building	5	0	2	13	4	2	219	6
Bedroom	16	28	29	61	15	40	58	192

3.2.2. Results

From each category, in the considered data set, we randomly chose 100 images for model training. Unlike text classification which usually leads to sparse training matrices, the abundance of visual descriptors in the images and the relatively lower number of extracted visual keywords, as compared to the textual vocabulary, lead to less sparse matrices for scenes classification. In Fig. 6, we compare the success rates of the LGDA and LDA models, when varying the number of extracted latent topics, $d+1$ in Section 2, over the data set. According to this figure it is clear that better categorization results are obtained when adopting LGDA. Fig. 7 shows examples of per class comparisons between the success rates obtained by both models. It is obvious that due to the less sparse nature of the count data vectors extracted in the case of scenes classification task (as compared to the text classification task presented in the previous section), the better fitting capabilities of the LGDA become more evident. Tables 4 and 5 show the optimal confusion matrices of the LGDA and LDA models. It should be noted that in few instances it happens that the likelihood of two or more classes become equal. This happens when the extracted visual keywords entirely fall outside the visual keywords present in the training set. In these rare cases the models are set to drop the scene altogether. According to these tables it is clear again that the LGDA gives significantly a better classification accuracy (65.69%) than the LDA (62.49%).

3.3. Comparison of the computational requirements of the LGDA versus the LDA

An essential concern when proposing new models as replacements for already established ones is the tradeoff between what the model offers and what it requires in return. LGDA in general is a more computationally demanding model than LDA. Indeed, in dimension d the Dirichlet has $d+1$ parameters while the generalized Dirichlet has $2d$ parameters. Thus, comparing to the Dirichlet, the generalized Dirichlet has $d-1$ extra parameters which is a very important advantage. Indeed, as the Dirichlet has $d+1$ parameters, when constructing a Dirichlet prior and if the mean probabilities of the variables have been fixed, it remains only one degree of freedom (by fixing the value of $\sum_{l=1}^{d+1}\alpha_l$) to adjust the distribution (Bouguila, 2008). For the generalized Dirichlet, however, it remains d degrees of freedom which makes it more flexible. Thus, for the same number of latent topics, the number of parameters related to the prior choice, that we

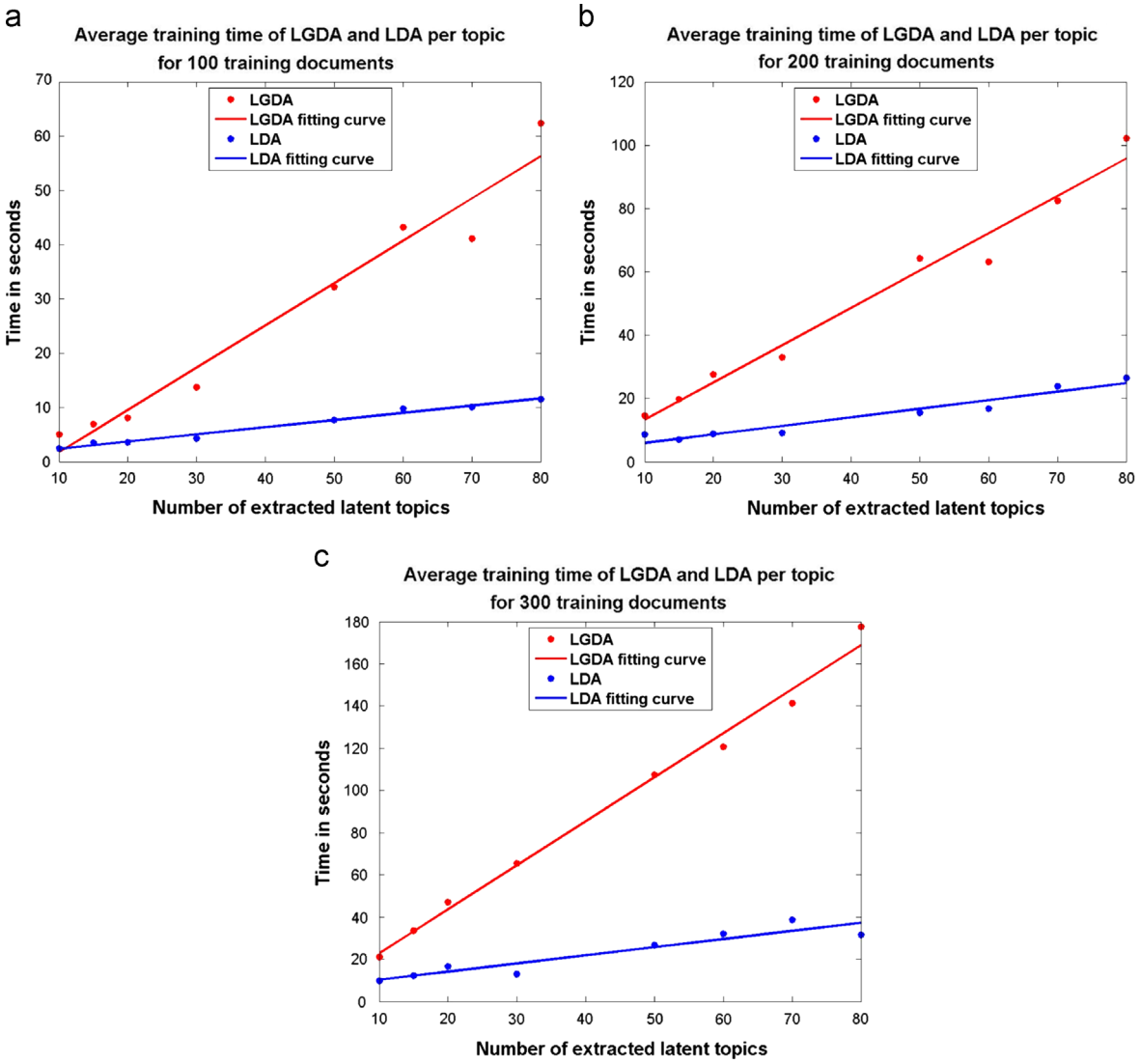

Fig. 8. Comparison of the computational time needed for training the LDA and LGDA models, for different numbers of training documents, as a function of the number of latent topics. The numbers of considered training documents are: (a) 100, (b) 200, and (c) 300. Red line: LGDA, blue Line: LDA. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

need to estimate via variational inference in the LGDA case, is almost twice the number needed for LDA. The other parameters remain the same. One concern is the computational requirements of the model parameters estimation regarding the inversion of the Hessian matrix in both models. It was shown in this paper that like the LDA case the computation of the Hessian matrix in LGDA is a linearly related to the number of generalized Dirichlet parameters. To show the computational requirements of our model in comparison with LDA we proceed with performing a series of experiments depicting the time it takes for both models to learn their parameters in different learning conditions. The result of these experiments is shown in Fig. 8. From this figure, we can see that although in general LGDA is a more computationally demanding model, like LDA, the computational demand for additional extracted topics clearly follows a linear curve.

4. Conclusion

In this work, an elegant generalized version of the LDA model has been developed. Unlike the earliest efforts that were directed to the extension of the LDA model (e.g. to online or hierarchical settings, for instance) by keeping its basic Dirichlet assumption, we propose an extension that offers greater generality by resorting to the generalized Dirichlet distribution. The recourse to the generalized Dirichlet is mainly motivated by the excellent results that we obtained recently by its adoption, as both prior and parent distribution, in several statistical modeling frameworks. The LGDA model has the chief advantage of containing the LDA model as a special case and then provides more versatile capabilities than the basic LDA model. We tested our proposed model on two challenging applications namely text and visual scenes classification. The obtained results demonstrate superior performance of the LGDA. The flexibility and generalization capabilities of our model offer vast future research opportunities in the different fields where the LDA model has been previously applied. Promising future works could be devoted to the extension of our model to handle hierarchical topic models as well as its adaptation for online learning settings. The potential of the LGDA model is overwhelming and it is our hope that it will serve to inspire more interesting applications and learning techniques.

Acknowledgment

The completion of this research was made possible thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC). The complete source code of this work is available upon request.

Appendix A. Exponential form of the generalized Dirichlet distribution

Here, we present the exponential form of the generalized Dirichlet distribution. The exponential form delivers us certain relationships necessary for developing the variational Bayes inference that we shall adopt. It is straightforward to show the generalized Dirichlet can be written in the following exponential form (Bouguila, 2012):

$$p(\vec{\theta}\mid\vec{\xi})=Z(\vec{\xi})\times\exp\left[\sum_{l=1}^{2d}G_l(\vec{\xi})T_l(\vec{\theta})\right] \tag{23}$$

In above we have

$$Z(\vec{\xi})=\prod_{l=1}^{d}\frac{\Gamma(\alpha_l+\beta_l)}{\Gamma(\alpha_l)\times\Gamma(\beta_l)}$$

$$G_l(\vec{\xi})=\alpha_l,\quad l=1,\ldots,d$$

$$G_l(\vec{\xi})=\beta_{l-d}-\alpha_{l-d+1}-\beta_{l-d+1},\quad l=d+1,\ldots,2d-1$$

$$G_{2d}(\vec{\xi})=\beta_d$$

$$T_l(\vec{\theta})=\log(\theta_l),\quad l=1,\ldots,d$$

$$T_l(\vec{\theta})=\log\left(1-\sum_{t=1}^{l-d}\theta_t\right),\quad l=d+1,\ldots,2d$$

In the above $Z(\vec{\xi})$ is the normalization factor, $\vec{G}(\vec{\xi})=(G_1(\vec{\xi}),\ldots,G_{2d}(\vec{\xi}))$ is the natural parameter and $\vec{T}(\vec{\theta})=(T_1(\vec{\theta}),\ldots,T_{2d}(\vec{\theta}))$ is the sufficient statistics of the distribution. For the exponential family of distributions, we know that the derivative of the logarithm of normalization factor with respect to the natural parameters equals the expected value of the sufficient statistics. Therefore, we have

$$E[\log(\theta_l)]=\Psi(\alpha_l+\beta_l)-\Psi(\alpha_l)-\Psi(\beta_l),\quad l=1,\ldots,d. \tag{24}$$

$$E\left[\log\left(1-\sum_{t=1}^{l}\theta_t\right)\right]=\Psi(\beta_l)-\Psi(\alpha_l+\beta_l),\quad l=1,\ldots,d. \tag{25}$$

Appendix B. Breakdown of $L(\vec{\xi}_q,\Phi_{\boldsymbol{w}};\vec{\xi},\mu_w)$

By factorizing $L(\vec{\xi}_q,\Phi_{\boldsymbol{w}};\vec{\xi},\mu_w)$ in Eq. (12), we obtain

$$\begin{aligned}L(\vec{\xi}_q,\Phi_{\boldsymbol{w}};\vec{\xi},\mu_w)=&E_q[\log\ p(\vec{\theta}\mid\vec{\xi})]+E_q[\log\ p(\boldsymbol{z})]\\&+E_q[\log\ p(\boldsymbol{w}|\boldsymbol{z},\mu_w)]-E_q[\log\ q(\vec{\theta})]-E_q[\log\ q(\boldsymbol{z})]\end{aligned} \tag{26}$$

We proceed with deriving each of the five factors of the above equation in the following:

$$\begin{aligned}E_q[\log\ p(\vec{\theta}\,|\xi)]=&\sum_{l=1}^{d}[\log\Gamma(\alpha_l+\beta_l)-\log\ \Gamma(\alpha_l)-\log\Gamma(\beta_l)]\\&+\sum_{l=1}^{d}[(\Psi(\gamma_l)-\Psi(\gamma_l+\delta_l))\alpha_l\\&+(\Psi(\delta_l)-\Psi(\gamma_l+\delta_l))(\beta_l-\alpha_{l+1}-\beta_{l+1}))]\end{aligned} \tag{27}$$

where $\gamma_l=\alpha_l+\sum_{n=1}^{N}\phi_{nl}$ and $\delta_l=\beta_l+\sum_{n=1}^{N}\sum_{ll=l+1}^{d+1}\phi_{n(ll)}$ (see Appendix B.2)

$$\begin{aligned}E_q(\log\ p(\boldsymbol{z}|\vec{\theta}))=&\sum_{n=1}^{N}\sum_{l=1}^{d}\phi_{nl}(\Psi(\gamma_l)-\Psi(\gamma_l+\delta_l))\\&+\sum_{n=1}^{N}\phi_{nd+1}(\Psi(\delta_d)-\Psi(\gamma_d+\delta_d))\end{aligned} \tag{28}$$

$$E_q[\log\ p(\boldsymbol{w}|\boldsymbol{z},\mu_w)]=\sum_{n=1}^{N}\sum_{l=1}^{d+1}\sum_{j=1}^{V}\phi_{nl}w_n^j\log(\mu_{w(lj)}) \tag{29}$$

where $\mu_{w(lj)}=p(w_n^j=1|z^l=1)$

$$\begin{aligned}E_q[\log\ q(\vec{\theta})]=&\sum_{l=1}^{d}(\log\ \Gamma(\gamma_l+\delta_l)-\log\ \ \Gamma(\gamma_l)-\log\ \ \Gamma(\delta_l))\\&+\sum_{l=1}^{d}[(\Psi(\gamma_l)-\Psi(\gamma_l+\delta_l))\gamma_l\\&+(\Psi(\delta_l)-\Psi(\gamma_l+\delta_l))(\delta_l-\gamma_{l+1}-\delta_{l+1})]\end{aligned} \tag{30}$$

$$E_q[\log\ q(\boldsymbol{z})]=\sum_{n=1}^{N}\sum_{l=1}^{d=1}\phi_{nl}\log(\phi_{nl}) \tag{31}$$

Having the above formulas we proceed with finding the parameters estimates.

B.1. Variational multinomial

In order to derive the parameter ϕ_{nl}, the probability that the n-th word is generated by the l-th hidden topic, we proceed with maximizing Eq. (26) with respect to ϕ_{nl}. Firstly, we separate the terms in Eq. (26) containing ϕ_{nl}:

$$L[\phi_{nl}] = \phi_{nl}(\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l)) + \phi_{nl} \log \mu_{w(lv)} - \phi_{nl} \log \phi_{nl} + \lambda_n \left(\sum_{ll=1}^{d+1} \phi_{n(ll)} - 1 \right) \tag{32}$$

and

$$L[\phi_{n(d+1)}] = \phi_{n(d+1)}(\Psi(\delta_d) - \Psi(\gamma_d + \delta_d)) + \phi_{n(d+1)} \log \beta_{(d+1)v} - \phi_{n(d+1)} \log \phi_{n(d+1)} + \lambda_n \left(\sum_{ll=1}^{d+1} \phi_{n(ll)} - 1 \right) \tag{33}$$

and therefore we have

$$\frac{\partial L}{\partial \phi_{nl}} = (\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l)) + \log \beta_{lv} - \log \phi_{nl} - 1 + \lambda_n \tag{34}$$

and

$$\frac{\partial L}{\partial \phi_{n(d+1)}} = (\Psi(\delta_d) - \Psi(\gamma_d + \delta_d)) + \log \beta_{(d+1)v} - \log \phi_{n(d+1)} - 1 + \lambda_n \tag{35}$$

Setting the above equation to zero leads to

$$\phi_{nl} = \beta_{lv} e^{(\lambda_n - 1)} e^{(\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l))} \tag{36}$$

$$\phi_{n(d+1)} = \beta_{(d+1)v} e^{(\lambda_n - 1)} e^{(\Psi(\delta_d) - \Psi(\gamma_d + \delta_d))} \tag{37}$$

considering that $\sum_{ll=1}^{d+1} \phi_{n(ll)} = 1$ for the normalization factor, we have

$$e^{\lambda_n - 1} = \frac{1}{\sum_{l=1}^{d} \beta_{lv} e^{(\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l))} + \beta_{(d+1)v} e^{(\Psi(\delta_d) - \Psi(\gamma_d + \delta_d))}}$$

B.2. Variational generalized Dirichlet

To find the updating equations for the variational generalized Dirichlet we again proceed with separating the terms in Eq. (26) containing the variational generalized Dirichlet parameters:

$$\begin{aligned} L[\vec{\xi}_q] = & \sum_{l=1}^{d} [(\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l))\alpha_l + (\Psi(\delta_l) - \Psi(\gamma_l + \delta_l)) \\ & \times (\beta_l - \alpha_{l+1} - \beta_{l+1})] + \sum_{n=1}^{N} \phi_{nl}(\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l)) \\ & + \sum_{n=1}^{N} \phi_{n(d+1)}(\Psi(\gamma_d) - \Psi(\gamma_d + \delta_d)) \\ & - \Bigg[\sum_{l=1}^{d} (\log \Gamma(\gamma_l + \delta_l) - \log \Gamma(\gamma_l) - \log \Gamma(\delta_l)) \\ & + \sum_{l=1}^{d} [(\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l))\gamma_l + (\Psi(\delta_l) \\ & - \Psi(\gamma_l + \delta_l))(\delta_l - \gamma_{l+1} - \delta_{l+1})] \Bigg] \end{aligned} \tag{38}$$

Setting the derivative of the above equation to zero leads to the following updating equations:

$$\gamma_l = \alpha_l + \sum_{n=1}^{N} \phi_{nl} \tag{39}$$

$$\delta_l = \beta_l + \sum_{n=1}^{N} \sum_{ll=l+1}^{d+1} \phi_{n(ll)} \tag{40}$$

B.3. Topic based multinomial

In this appendix we derive the updating equations necessary for estimating μ_w. Maximizing Eq. (26) with respect to μ_w leads to the same equation as in the LDA case:

$$L[\mu_w] = \sum_{d=1}^{M} \sum_{n=1}^{N_s} \sum_{l=1}^{k+1} \sum_{j=1}^{V} \phi_{dnl} w_{dn}^{j} \log \mu_{w(lj)} + \sum_{l=1}^{k+1} \lambda_l \left(\sum_{j=1}^{V} \mu_{w(ij)} - 1 \right) \tag{41}$$

Taking the derivative with respect to $\mu_{w(lj)}$ and setting it to zero gives

$$\mu_{w(lj)} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dnl} w_{dn}^{j} \tag{42}$$

B.4. Generalized Dirichlet parameters

We choose the terms of Eq. (26) containing the generalized Dirichlet parameters $\vec{\xi}$:

$$\begin{aligned} L[\vec{\xi}] = & \sum_{m=1}^{M} (\log(\Gamma(\alpha_l + \beta_l)) - \log(\Gamma(\alpha_l)) - \log(\Gamma(\beta_l))) \\ & + \sum_{m=1}^{M} [(\Psi(\gamma_{ml}) - \Psi(\gamma_{ml} + \delta_{ml}))\alpha_l \\ & + (\Psi(\delta_{ml}) - \Psi(\gamma_{ml} + \delta_{ml}))\beta_l] \end{aligned} \tag{43}$$

The derivative of the above with respect to the generalized Dirichlet parameters gives

$$\frac{\partial L[\vec{\xi}]}{\partial \alpha_l} = M(\Psi(\alpha_l + \beta_l) - \Psi(\alpha_l)) + \sum_{m=1}^{M} (\Psi(\gamma_{ml}) - \Psi(\gamma_{ml} + \delta_{ml})) \tag{44}$$

and

$$\frac{\partial L[\vec{\xi}]}{\partial \beta_l} = M(\Psi(\alpha_l + \beta_l) - \Psi(\beta_l)) + \sum_{m=1}^{M} (\Psi(\delta_{ml}) - \Psi(\gamma_{ml} + \delta_{ml})) \tag{45}$$

It can be seen from the two previous equations that the derivative of the lower bound (Eq. (26)) with respect to each of the generalized Dirichlet parameters α_l and β_l depend not only on their own values, but also on each other. To solve the optimization problem we use the Newton–Raphson method. Thus, we need to compute the Hessian matrix which takes in our case a peculiarly interesting form:

$$\frac{\partial^2 L[\vec{\xi}]}{\partial \alpha_l^2} = M(\Psi'(\alpha_l + \beta_l) - \Psi'(\alpha_l)) \tag{46}$$

$$\frac{\partial^2 L[\vec{\xi}]}{\partial \beta_l^2} = M(\Psi'(\alpha_l + \beta_l) - \Psi'(\beta_l)) \tag{47}$$

$$\frac{\partial^2 L[\vec{\xi}]}{\partial \alpha_l \partial \beta_l} = M(\Psi'(\alpha_l + \beta_l)) \tag{48}$$

$$\frac{\partial^2 L[\vec{\xi}]}{\partial \beta_l \partial \alpha_l} = M(\Psi'(\alpha_l + \beta_l)) \tag{49}$$

The other entries of the Hessian matrix are zeros. According to the previous four equations the Hessian matrix has a block diagonal form and therefore its inverse will be the inverse of 2×2 matrices on the diagonal which can be easily computed.

References

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003a. Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Blei, D.M., Griffiths, T.L., Jordan, M.I., Tenenbaum, J.B., 2003b. Hierarchical topic models and the nested Chinese restaurant process. In: Advances in Neural Information Processing Systems (NIPS), MIT Press.
Bouguila, N., 2008. Clustering of count data using generalized Dirichlet multinomial distributions. IEEE Trans. Knowl. Data Eng. 20, 462–474.
Bouguila, N., 2011. Count data modeling and classification using finite mixtures of distributions. IEEE Trans. Neural Netw. 22, 186–198.
Bouguila, N., 2012. Hybrid generative/discriminative approaches for proportional data modeling and classification. IEEE Trans. Knowl. Data Eng. 24, 2184–2202.
Bouguila, N., Ziou, D., 2007a. Unsupervised learning of a finite discrete mixture: applications to texture modeling and image databases summarization. J. Vis. Commun. Image Represent. 18, 295–309.
Bouguila, N., Ziou, D., 2007b. High-dimensional unsupervised selection and estimation of a finite generalized Dirichlet mixture model based on minimum message length. IEEE Trans. Pattern Anal. Mach. Intell. 29, 1716–1731.
Bouguila, N., Ziou, D., 2009. A nonparametric Bayesian learning model: application to text and image categorization. In: Theeramunkong, T., Kijsirikul, B., Cercone, N., Ho, T.B. (Eds.), PAKDD, Lecture Notes in Computer Science, vol. 5476. Springer, pp. 463–474.
Boutemedjet, S., Ziou, D., Bouguila, N., 2007. Unsupervised feature selection for accurate recommendation of high-dimensional image data. In: Advances in Neural Information Processing Systems (NIPS), MIT Press, pp. 177–184.
Caballero, K.L., Barajas, J., Akella, R., 2012. The generalized Dirichlet distribution in enhanced topic detection. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM), ACM, pp. 773–782.
Carneiro, G., Chan, A.B., Moreno, P.J., Vasconcelos, N., 2007. Supervised learning of semantic classes for image annotation and retrieval. IEEE Trans. Pattern Anal. Mach. Intell. 29, 394–410.
Connor, R.J., Mosimann, J.E., 1969. Concepts of independence for proportions with a generalization of the Dirichlet distribution. J. Am. Stat. Assoc. 64, 194–206.
Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C., 2004. Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, 8th European Conference on Computer Vision (ECCV), Springer, pp. 1–12.
Deerwester, S., Dumais, S., Landauer, T., Furnas, G., Harshman, R., 1990. Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. 41, 391–407.
Fan, W., Bouguila, N., Ziou, D., 2012. Variational learning for finite Dirichlet mixture models and applications. IEEE Trans. Neural Netw. Learn. Syst. 23, 762–774.
Fei-Fei, L., Perona, P., 2005. A Bayesian hierarchical model for learning natural scene categories. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Press, pp. 524–531.
Hoffman, M.D., Blei, D.M., Bach, F.R., 2010. Online learning for latent Dirichlet allocation. In: Advances in Neural Information Processing Systems (NIPS), Curran Associates, Inc., pp. 856–864.
Hofmann, T., 2001. Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. 42, 177–196.
Joachims, T., 1998. Text categorization with support vector machines: learning with many relevant features. In: Proceedings of the 10th European Conference on Machine Learning (ECML), Springer, pp. 137–142.
Jordan, M.I., Ghahramani, Z., Jaakkola, T., Saul, L.K., 1999. An introduction to variational methods for graphical models. Mach. Learn. 37, 183–233.
Lacoste-Julien, S., Sha, F., Jordan, M.I., 2008. DiscLDA: discriminative learning for dimensionality reduction and classification. In: Advances in Neural Information Processing Systems (NIPS), Curran Associates, Inc., pp. 897–904.
Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60, 91–110.
Madsen, R.E., Kauchak, D., Elkan, C., 2005. Modeling word burstiness using the Dirichlet distribution. In: Proceedings of the 22nd International Conference on Machine Learning (ICML). ACM Press, Bonn, Germany, pp. 545–552.
Mei, Q., Fang, H., Zhai, C., 2007. A study of poisson query generation model for information retrieval. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information (SIGIR). ACM, pp. 319–326.
Minka, T.P., 2007. Estimating a Dirichlet distribution. Unpublished paper available at: ⟨http://research.microsoft.com/∼minka/papers/dirichlet/⟩. Microsoft Research, Cambridge, UK.
Nigam, K., McCallum, A., Thrun, S., Mitchell, T., 2000. Text classification from labeled and unlabeled documents using EM. Mach. Learn. 39, 103–134.
Oliva, A., Torralba, A., 2001. Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vis. 42, 145–175.
Rashedi, E., Nezamabadi-pour, H., Saryazdi, S., 2013. A simultaneous feature adaptation and feature selection method for content-based image retrieval systems. Knowl.-Based Syst. 39, 85–94.
Robert, C., Casella, G., 2004. Monte Carlo Statistical Methods, second ed. Springer.
Ruiz, M.E., Srinivasan, P., 2002. Hierarchical text categorization using neural networks. Inf. Retr. 5, 87–118.
Sebastiani, F., 2002. Machine learning in automated text categorization. ACM Comput. Surv. 34, 1–47.
Sivic, J., Russell, B.C., Zisserman, A., Freeman, W.T., Efros, A.A., 2008. Unsupervised discovery of visual object class hierarchies. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Press, pp. 1–8.
Stokes, N., Carthy, J., 2001. Combining semantic and syntactic document classifiers to improve first story detection. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). ACM, pp. 424–425.
van Gemert, J., Veenman, C.J., Smeulders, A.W.M., Geusebroek, J.-M., 2010. Visual word ambiguity. IEEE Trans. Pattern Anal. Mach. Intell. 32, 1271–1283.
Vinokourov, A., Girolami, M., 2002. A probabilistic framework for the hierarchic organisation and classification of document collections. J. Intell. Inf. Syst. 18, 153–172.
Watanabe, K., Watanabe, S., 2006. Stochastic complexities of Gaussian mixtures in variational Bayesian approximation. J. Mach. Learn. Res. 7, 625–644.
Zhang, J., Chen, L., Guo, G., 2013. Projected-prototype based classifier for text categorization. Knowl.-Based Syst. 49, 179–189.