Online Learning for Two Novel Latent Topic Models

Ali Shojaee Bakhtiari, Nizar Bouguila

Online Learning for Two Novel Latent Topic Models

Ali Shojaee Bakhtiari[1] and Nizar Bouguila[2]

[1] Department of Electrical and Computer engineering
Concordia University, Montreal, QC, Canada
al_sho@encs.concordia.ca
[2] Concordia Institute for Information Systems Engineering
Concordia University, Montreal, QC, Canada
nizar.bouguila@concordia.ca

Abstract. Latent topic models have proven to be an efficient tool for modeling multitopic count data. One of the most well-known models is the latent Dirichlet allocation (LDA). In this paper we propose two improvements for LDA using generalized Dirichlet and Beta-Liouville prior assumptions. Moreover, we apply an online learning approach for both introduced approaches. We choose a challenging application namely natural scene classification for comparison and evaluation purposes.

Keywords: Generalized Dirichlet, Beta-Liouville, online learning , Latent model, variational learning, count data.

1 Introduction

In order to extract the hidden information within count data various models have been proposed in the past. The first widely used model for count data modeling was the naive Bayes model combined with multinomial distribution [14, 6, 1]. However several researchers proceeded with mentioning the oversimplifications and the subsequent drawbacks of the Naive Bayes assumption [13, 5, 8]. The foremost solution offered to compensate for the deficiencies of the naive assumption was considering the Dirichlet distribution [13, 5] as the prior assumption for the multinomial distribution. Based on the Dirichlet assumption, several models have been developed for proper count data modeling. One model that has gained much acceptance among the research community is the latent Dirichlet allocation (LDA) model firstly proposed in [2]. LDA model uses a Bayesian model for data generation using a variational Bayes (VB) approach for parameter inference. The majority of the models developed thus far have been based on the Dirichlet prior assumption. However, researchers began questioning the merit of Dirichlet assumption [7]. The main drawback of the Dirichlet assumption is the fact that it has a strictly negative covariance matrix and therefore it inherently fails to properly model the data in which topics have positive correlation in between. The other main drawback of the Dirichlet distribution is the fact that the elements with similar mean need to have similar variance, which

clearly is an oversimplification. To overcome these shortcomings research has recently been shifted towards finding models with better modeling accuracy.
Recently it has been shown that the generalized Dirichlet distribution is a good replacement for the Dirichlet distribution when using finite mixture models [3]. One important factor of the generalized Dirichlet is that like Dirichlet distribution it is a conjugate prior to the multinomial distribution. Generalized Dirichlet distribution also does not carry the restrictions of the Dirichlet distribution and allows more relaxed modeling capabilities. Another modeling prior that has recently attracted notice is the Beta-Liouville distribution [7]. The advantage point of the Beta-Liouville distribution in comparison to generalized Dirichlet is that it requires only two more parameters to be estimated compared to Dirichlet distribution compared with the twice the parameters the generalized Dirichlet distribution requires. Based on the above facts we proceed with proposing two different latent topic models based on the LDA model but with the generalized Dirichlet and the Beta-Liouville assumptions. The former model is called latent generalized Dirichlet allocation (LGDA) and the latter is called latent Beta-Liouville allocation (LBLA). The learning of both models is performed online using the approach proposed in [10].
The structure of the paper is as follows. In section 2, we introduce the LGDA and LBLA models. In section 3 we shall describe the adaption of the online learning model on the two models. In section 4 we shall bring the experimental results and we will finalize the paper with conclusion.

2 Proposed Latent Topic Models

In this section we briefly describe the two proposed latent topic models. Both models essentially have the same generative model as the LDA:

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

In above z_n is a d dimensional binary vector of topics defined so that $z_n^i = 1$ if the $i-th$ topic is chosen and zero, otherwise. We define, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$. A chosen topic is attributed to a multinomial prior β_w over the vocabulary of words so that $\beta_{w_{(ij)}} = p(w^j = 1|z^i = 1)$, from which every word is randomly drawn. $p(w_n|z_n, \beta_w)$ is a multinomial probability conditioned on z_n and $Dir(\boldsymbol{\xi})$ is a d-variate Dirichlet distribution with parameters $\boldsymbol{\xi} = (\alpha_1, \ldots, \alpha_d)$.
The main inference problem of LGDA is estimating the posterior of the hidden variables, $\boldsymbol{\theta}$ and $\boldsymbol{z}$:

$$p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\xi}, \beta_w) = \frac{p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}|\boldsymbol{\xi}, \beta_w)}{p(\boldsymbol{w}|\boldsymbol{\xi}, \beta_w)} \tag{1}$$

The above equation is known to be intractable. As proposed in [2], an efficient way to estimate the parameters of this intractable posterior is to use the vibrational 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 vibrational distributions on the hidden variables [11]:

$$q(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\xi_q}, \Phi_{\boldsymbol{w}}) = q(\boldsymbol{\theta}|\boldsymbol{\xi_q}) \prod_{n=1}^{N} q(z_n|\phi_n) \tag{2}$$

The details of the parameter estimation algorithm can de found in [2]. In this section we proceed with introducing LGDA and LBLA subsequently.

2.1 Latent Generalized Dirichlet Allocation

The major difference between the LGDA and the LDA model is the consideration of the generalized Dirichlet assumption:

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

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, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{d+1})$, where $\theta_{d+1} = 1 - \sum_{i=1}^{d} \theta_i$. A chosen topic is attributed to a multinomial prior β_w over the vocabulary of words so that $\beta_{w_{(ij)}} = p(w^j = 1|z^i = 1)$, from which every word is randomly drawn. $p(w_n|z_n, \beta_w)$ is a multinomial probability conditioned on z_n and $GenDir(\boldsymbol{\xi})$ is a d-variate generalized Dirichlet distribution with parameters $\boldsymbol{\xi} = (\alpha_1, \beta_1, \ldots, \alpha_d, \beta_d)$ and probability distribution function given by:

$$p(\theta_1, \ldots, \theta_d|\boldsymbol{\xi}) = \prod_{i=1}^{d} \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} \theta_i^{\alpha_i - 1} (1 - \sum_{j=1}^{i} \theta_j)^{\gamma_i} \tag{3}$$

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 [4]. Therefore it is understood that under certain conditions LGDA will also behave like LDA and subsequently LDA is a special case of the LGDA model. To estimate the posterior of the hidden variables of the LGDA model we use a similar VB approach as the one proposed for LDA, where $q(\boldsymbol{\theta}|\boldsymbol{\xi_q})$ 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 $\boldsymbol{\Phi_w} = \{\phi_1, \phi_2, \ldots, \phi_N\}$. Using Jensen's inequality [11] one can derive the following:

$$\log p(\boldsymbol{w}|\boldsymbol{\xi}, \beta_w) \geq E_q[\log p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}|\boldsymbol{\xi}, \beta_w)] - E_q[\log q(\boldsymbol{\theta}, \boldsymbol{z})] \tag{4}$$

Assigning $L(\boldsymbol{\xi_q}, \boldsymbol{\Phi_w}; \boldsymbol{\xi}, \beta_w)$ 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:

$$\log p(\boldsymbol{w}|\boldsymbol{\xi}, \beta_w) = L(\boldsymbol{\xi_q}, \boldsymbol{\Phi_w}; \boldsymbol{\xi}, \beta_w) + KL\big(q(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{\xi_q}, \boldsymbol{\Phi_w}||p(\boldsymbol{\theta}, \boldsymbol{z})|\boldsymbol{w}, \boldsymbol{\xi}, \beta_w)\big) \quad (5)$$

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 $L(\boldsymbol{\xi_q}, \boldsymbol{\Phi_w}; \boldsymbol{\xi}, \beta_w)$. 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. The breakdown of $L(\boldsymbol{\xi_q}, \boldsymbol{\Phi_w}; \boldsymbol{\xi}, \beta_w)$ to maximize the lower bound $L(\boldsymbol{\xi_q}, \boldsymbol{\Phi_w}; \boldsymbol{\xi}, \beta_w)$ with respect to ϕ_{nl}, leads to the following updating equations for the variational multinomial:

$$\phi_{nl} = \beta_{lv} e^{(\lambda_n - 1)} e^{(\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l))} \quad \phi_{n(d+1)} = \beta_{(d+1)v} e^{(\lambda_n - 1)} e^{(\Psi(\delta_d) - \Psi(\gamma_d + \delta_d))} \quad (6)$$

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:

$$e^{\lambda_n - 1} = \frac{1}{\sum_{l=1}^{d} \beta_{lv} e^{(\Psi(\gamma_l) - \Psi(\gamma_l + \delta_l))} + \beta_{(d+1)v} e^{(\Psi(\delta_d) - \Psi(\gamma_d + \delta_d))}} \quad (7)$$

Maximizing the lower bound L with respect to the variational generalized Dirichlet parameter gives the following updating equations:

$$\gamma_l = \alpha_l + \sum_{n=1}^{N} \phi_{nl} \quad \delta_l = \beta_l + \sum_{n=1}^{N} \sum_{ll=l+1}^{d+1} \phi_{n(ll)} \quad (8)$$

The above equations show 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 [2] for the LDA case. This is a direct result of the conjugacy between the generalized Dirichlet and the multinomial distribution. 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 $\beta_{w(lj)}$ delivers the following updating equation:

$$\beta_{w(lj)} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dnl} w_{dn}^{j} \quad (9)$$

The model's parameters are derived using a Newton-Raphson method.

2.2 Latent Beta-Liouville allocation

The model that we briefly discuss in this subsection, latent Beta-Liouville allocation (LBLA), is another model developed based on the LDA model. The model assumes a Beta-Liouville as its topic generating prior. 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 $(\theta_1, \ldots, \theta_D) \propto BL(\boldsymbol{\xi})$.
3. For each of the N words w_n:
   (a) choose a topic $z_n \propto Multinomial(\boldsymbol{\theta})$.
   (b) Choose a word w_n from $p(w_n|z_n, \beta_w)$.

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, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{D+1})$, where $\theta_{D+1} = 1 - \sum_{i=1}^{D} \theta_i$. A chosen topic is attributed to a multinomial prior β_w over the vocabulary of words so that $\beta_{w_{(ij)}} = p(w^j = 1|z^i = 1)$, from which every word is randomly drawn. $p(w_n|z_n, \beta_w)$ is a multinomial probability conditioned on z_n and $BL(\boldsymbol{\xi})$ is a d-variate Beta-Liouville distribution with parameters $\boldsymbol{\xi} = (\alpha_1, \beta_1, \ldots, \alpha_d, \beta_d)$ and probability distribution function given by:

$$P(\theta_1, \ldots, \theta_D|\boldsymbol{\xi}) = \frac{\Gamma(\sum_{d=1}^{D} \alpha_d)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \prod_{d=1}^{D} \frac{\theta_d^{\alpha_d - 1}}{\Gamma(\alpha_d)} (\sum_{d=1}^{D} \theta_d)^{\alpha - \sum_{l=1}^{D} \alpha_l} (1 - \sum_{l=1}^{D} \theta_l)^{\beta - 1} \tag{10}$$

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 Beta-Liouville distribution is reduced to Dirichlet distribution [4]. We define, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{D+1})$, where $\theta_{D+1} = 1 - \sum_{i=1}^{D} \theta_i$. The maximization of the lower bound $L(\boldsymbol{\xi_q}, \boldsymbol{\Phi_w}; \boldsymbol{\xi}, \beta_w)$ with respect to ϕ_{nl}, leads to the following updating equations:

$$\phi_{nl} = \beta_{lv} e^{(\lambda_n - 1)} e^{(\Psi(\gamma_i) - \Psi(\sum_{ii=1}^{D} \gamma_{ii}))} \tag{11}$$

$$\phi_{n(D+1)} = \beta_{(D+1)v} e^{(\lambda_n - 1)} e^{(\Psi(\beta_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma))} \tag{12}$$

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:

$$e^{\lambda_n - 1} = \frac{1}{\beta_{(D+1)v} e^{(\Psi(\beta_\gamma) - \Psi(\alpha_\gamma + \beta_\gamma))} + \sum_{i=1}^{D} \beta_{iv} e^{(\Psi(\gamma_d) - \Psi(\sum_{ii=1}^{D} \gamma_{II}))}} \tag{13}$$

$$\gamma_i = \alpha + \sum_{n=1}^{N} \phi_{ni} \quad \alpha_\gamma = \alpha + \sum_{n=1}^{N} \sum_{d=1}^{D} \phi_{nd} \quad \beta_\gamma = \beta + \sum_{n=1}^{N} \phi_{n(D+1)} \tag{14}$$

The above equations show that the variational Beta-Liouville for each document acts as a posterior in the presence of the variational multinomial parameters.

The same conclusion was observed in [2] for the LDA case. This is a direct result of the conjugacy between the Beta-Liouville and the multinomial distribution. One needs to consider that the LBLA 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 $\beta_{w(lj)}$ delivers the following updating equation:

$$\beta_{w(lj)} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dnl} w_{dn}^{j} \tag{15}$$

The model's parameters are the last ones to be derived using a Newton-Raphson algorithm, also.

3 Online Latent Topic Models

The variational Bayes model of the LDA model and the subsequent adaption for the LGDA and LBLA are shown to coverage to a local likelihood of the actual posterior of the hidden parameters of the models. However, the main problem with the original VB approach is that it needs to consider the entire corpus beforehand for parameter estimation. This in return emerges two serious problems. Firstly, the need for the collection of the entire training corpus and secondly the computational requirements of dealing with a huge corpus. To overcome this problem the authors in [10] offered an online learning model that fixes the mentioned issues. The solution is based on a time dependent (time defined as the index of the part of the data given to the model in each iteration) weight:

$$\rho_t \triangleq (\tau_0 + t)^{-\kappa}, \kappa \in (0.5, 1] \tag{16}$$

The parameter τ_0 slows down the effect of early parameter estimations. The online learning algorithm can easily be extended to cover LGDA and LBLA models as well. The steps of the algorithm are as follows.

1. In each learning interval the model performs a batch VB over the patch of the training set attributed to that interval and assigns a weight value to the patch according to 16.
2. Prior parameter estimation: Perform the Newton-Raphson algorithm over the entire corpus for $t = 0$ to ∞ as: $\xi \leftarrow \xi - \rho_t \widetilde{\alpha}(\xi_t)$ where $\widetilde{\alpha}(\xi_t)$ is the inverse of the Hessian times the gradient in respect to α of the posterior lower bound.
3. Word dictionary update: $\widetilde{\beta_w}(t+1) = normalize((1-\rho_t)\widetilde{\beta_{wt}} + \rho_t \beta_w(t))$ where $\widetilde{\beta_w}(t)$ is the available estimation of the word dictionary at t-th step.

It was shown in [10] that the condition $\kappa \in (0.5, 1]$ is necessary for keeping the online learning model stable.

4 Experimental Results

In this section we shall proceed with applying our proposed two models, online LBLA and LGDA, on the challenging task of natural scene classification and make a comparison between the classification success rates offered by the two models versus that of the online LDA. The main idea that we use here is based on the description of scenes using visual words [9]. 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 [9, 17, 16].
For the construction of the visual words vocabulary, we need first to extract local descriptors from a set of training images. Many descriptors have been proposed in the past, but scale invariant feature transform (SIFT) descriptor [12], that we consider here, has dominated the literature. 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 [15]. The 7 classes chosen from the data set described in [15] are coast, forest, highway, inside of cities, open country, street, and tall building, which contain 361, 329, 261, 309, 411, 293, and 356 images, respectively. Examples of images from the different considered classes are shown in figure 1.

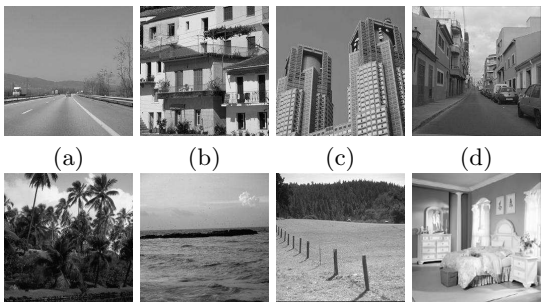

Fig. 1. Sample images from each group. (a) Highway, (b) Inside of cities, (c) Tall building, (d) Streets, (e) Forest, (f) Coast, (g) Open country, (h) Bedroom.

4.1 Comparison between the performance of LBLA and LGDA models against LDA

At first the models were given 5 chunks of training images each containing 20 images. In this set of experiments the effect of the online learning was reduced

since the small number of iterations plus the big chunks of test data quite resembled the Batch LDA and LGDA models. The results of applying the online LDA and LBLA models are brought in Fig. 2. Under the same experimental conditions we proceed with delivering the results for the LGDA model as well in Fig. 3. The optimal confusion matrix of the online LBLA model is brought

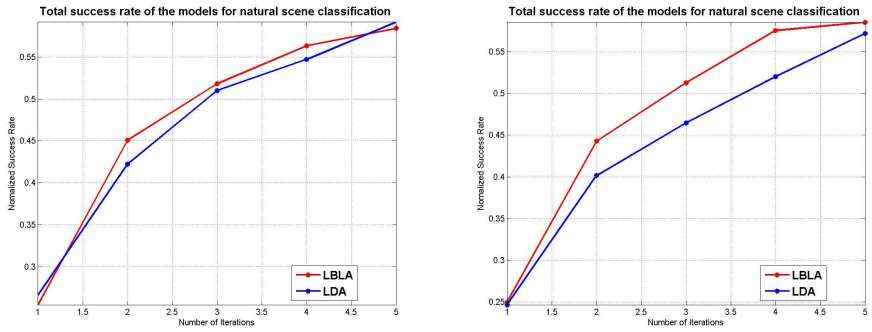

Fig. 2. Comparison of the success rates of the online LBLA model against online LDA model for the natural scene classification application for a training size that equals 20 for two different extracted number of topics.

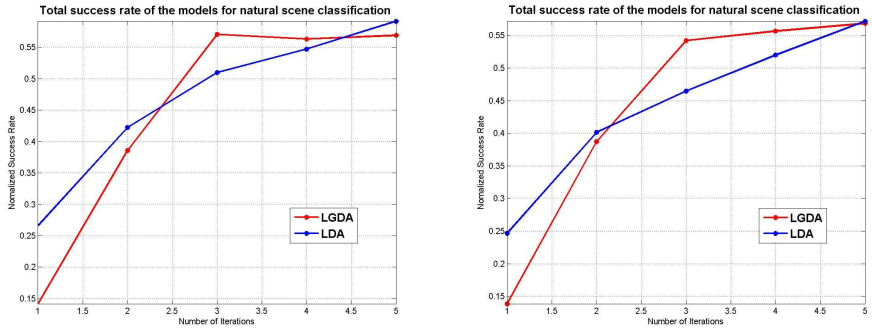

Fig. 3. Comparison of the success rate of the online LGDA model against online LDA model for the natural scene classification application over a training size that equals 20 for two different extracted number of topics.

in table 1. The optimal confusion matrix of the online LBLA model is brought in table 2 and the optimal confusion matrix of the online LDA model is brought in table 3.

	C	F	H	I	O	S	T
Coast (C)	216	1	86	3	50	3	2
Forest (F)	3	242	15	53	4	51	0
Highway (H)	40	1	69	5	6	3	15
Inside of cities (I)	0	2	4	146	2	12	6
Open country (O)	90	39	23	22	331	14	50
Streets (S)	5	41	60	57	9	203	2
Tall building (T)	6	2	3	22	8	6	281

Table 1. Optimal confusion matrix of the online LBLA model applied for the scenes classification task.

	C	F	H	I	O	S	T
Coast (C)	282	4	130	5	98	7	14
Forest (F)	9	287	24	84	24	167	3
Highway (H)	19	2	55	13	16	6	94
Inside of cities (I)	4	19	23	157	1	58	9
Open country (O)	38	7	19	24	241	19	66
Streets (S)	8	7	8	18	30	32	13
Tall building (T)	0	2	1	7	0	3	157

Table 2. Optimal confusion matrix of the online LGDA model applied for the scenes classification task.

	C	F	H	I	O	S	T
Coast (C)	316	25	133	40	296	45	175
Forest (F)	1	213	9	25	0	61	0
Highway (H)	1	0	15	0	0	0	0
Inside of cities (I)	2	35	18	187	1	62	14
Open country (O)	31	36	46	27	113	45	14
Streets (S)	0	11	28	3	0	69	0
Tall building (T)	9	8	11	26	0	10	153

Table 3. Optimal confusion matrix of the online LDA model applied for the scenes classification task.

5 Conclusion

In this work we proposed two new online learning multitopic models. We performed a series of experiments over a challenging application, natural scene classification, and we showed and compared the merits of our proposed models in comparison with online LDA. The two models show promising results when adapted for the online learning scheme and tend to surpass the online LDA model in the scope of the experiments performed.

Acknowledgments. The completion of this research was made possible thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Bakhtiari, A.S., Bouguila, N.: A novel hierarchical statistical model for count data modeling and its application in image classification. In: Huang, T., Zeng, Z., Li, C., Leung, C.S. (eds.) ICONIP (2). Lecture Notes in Computer Science, vol. 7664, pp. 332–340. Springer (2012)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
3. Bouguila, N.: Clustering of count data using generalized dirichlet multinomial distributions. IEEE Trans. Knowl. Data Eng. 20(4), 462–474 (2008)
4. Bouguila, N., Ziou, D.: High-dimensional unsupervised selection and estimation of a finite generalized Dirichlet mixture model based on minimum message length. IEEE Trans. on Pattern Analysis and Machine Intelligence 29(10), 1716–1731 (2007)
5. Bouguila, N., Ziou, D.: Unsupervised learning of a finite discrete mixture: Applications to texture modeling and image databases summarization. Journal of Visual Communication and Image Representation 18(4), 295–309 (2007)
6. Bouguila, N.: A model-based approach for discrete data clustering and feature weighting using map and stochastic complexity. IEEE Trans. Knowl. Data Eng. 21(12), 1649–1664 (2009)
7. Bouguila, N.: Count data modeling and classification using finite mixtures of distributions. IEEE Trans. on Neural Networks 22(2), 186–198 (2011)
8. Bouguila, N., ElGuebaly, W.: A generative model for spatial color image databases categorization. In: Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. 821–824 (2008)
9. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, 8th European Conference on Computer Vision (ECCV). pp. 1–12. Springer (2004)
10. Hoffman, M.D., Blei, D.M., Bach, F.R.: Online learning for latent dirichlet allocation. In: NIPS. pp. 856–864 (2010)
11. Jordan, M.I., Ghahramani, Z., Jaakkola, T., Saul, L.K.: An introduction to variational methods for graphical models. Machine Learning 37(2), 183–233 (1999)
12. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
13. Madsen, R.E., Kauchak, D., Elkan, C.: Modeling word burstiness using the Dirichlet distribution. In: Proc. of the 22^{nd} International Conference on Machine Learning (ICML). pp. 545–552. ACM Press, Bonn, Germany (2005)
14. Nigam, K., McCallum, A., Thrun, S., Mitchell, T.: Text classification from labeled and unlabeled documents using EM. Machine Learning 39(2), 103–134 (2000)
15. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 42(3), 145–175 (2001)
16. Scalzo, F., Piater, J.: Adaptive patch features for object class recognition with learned hierarchical models. In: Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–8. IEEE (2007)
17. Weston, J., Bengio, S., Usunier, N.: Large scale image annotation: learning to rank with joint word-image embeddings. Machine Learning 81(1), 21–35 (2010)