Journal of Machine Learning Research 13 (2012) 2233-2236 Submitted 10/11; Revised 1/12; Published 8/12

A Topic Modeling Toolbox Using Belief Propagation

Jia Zeng J.ZENG@IEEE.ORG
School of Computer Science and Technology
Soochow University
Suzhou 215006, China

Editor: Cheng Soon Ong

Abstract

Latent Dirichlet allocation (LDA) is an important hierarchical Bayesian model for probabilistic topic modeling, which attracts worldwide interests and touches on many important applications in text mining, computer vision and computational biology. This paper introduces a topic modeling toolbox (TMBP) based on the belief propagation (BP) algorithms. TMBP toolbox is implemented by MEX C++/Matlab/Octave for either Windows 7 or Linux. Compared with existing topic modeling packages, the novelty of this toolbox lies in the BP algorithms for learning LDA-based topic models. The current version includes BP algorithms for latent Dirichlet allocation (LDA), author-topic models (ATM), relational topic models (RTM), and labeled LDA (LaLDA). This toolbox is an ongoing project and more BP-based algorithms for various topic models will be added in the near future. Interested users may also extend BP algorithms for learning more complicated topic models. The source codes are freely available under the GNU General Public Licence, Version 1.0 at `https://mloss.org/software/view/399/`.

Keywords: topic models, belief propagation, variational Bayes, Gibbs sampling

1. Introduction

The past decade has seen rapid development of latent Dirichlet allocation (LDA) (Blei et al., 2003) for solving topic modeling problems because of its elegant three-layer graphical representation as well as two efficient approximate inference methods such as Variational Bayes (VB) (Blei et al., 2003) and collapsed Gibbs Sampling (GS) (Griffiths and Steyvers, 2004). Both VB and GS have been widely used to learn variants of LDA-based topic models until our recent work (Zeng et al., 2011) reveals that there is yet another learning algorithm for LDA based on loopy belief propagation (BP). The basic idea of BP is inspired by the collapsed GS algorithm, in which the three-layer LDA can be interpreted as being collapsed into a two-layer factor graph (Kschischang et al., 2001). The sum-product BP algorithm operates on the factor graph (Bishop, 2006). Extensive experiments confirm that BP is faster and more accurate than both VB and GS, and thus is a strong candidate for becoming the standard topic modeling algorithm. For example, we show how to learn three typical variants of LDA-based topic models, such as author-topic models (ATM) (Rosen-Zvi et al., 2004), relational topic models (RTM) (Chang and Blei, 2010), and labeled LDA (LaLDA) (Ramage et al., 2009) using BP based on the novel factor graph representations (Zeng et al., 2011).

We have implemented the topic modeling toolbox called TMBP by MEX C++ in the Matlab/Octave interface based on VB, GS and BP algorithms. Compared with other topic modeling packages,[1234567] the novelty of this toolbox lies in the BP algorithms for topic modeling. This paper introduces how to use this toolbox for basic topic modeling tasks.

2. Belief Propagation for Topic Modeling

Given a document-word matrix $\mathbf{x} = \{x_{w,d}\}$ ($x_{w,d}$ is the number of word counts at the index $\{w,d\}$) with word indices $1 \le w \le W$ in the vocabulary and document indices $1 \le d \le D$ in the corpus, the probabilistic topic modeling task is to allocate topic labels $\mathbf{z} = \{z^k_{w,d}\}, z^k_{w,d} \in \{0,1\}, \sum_{k=1}^K z^k_{w,d} = 1, 1 \le k \le K$ to partition the nonzero elements $x_{w,d} \neq 0$ into K topics (provided by the user) according to three topic modeling rules:

1. Co-occurrence: the different word indices w in the same document d tend to have the same topic label.
2. Smoothness: the same word indices w in the different documents d tend to have the same topic label.
3. Clustering: all word indices w do not tend to be associated with the same topic label.

Based on the above rules, recent approximate inference methods compute the marginal distribution of topic label $\mu_{w,d}(k) = p(z^k_{w,d} = 1)$ called *message*, and estimate parameters using the iterative EM (Bishop, 2006) algorithm according to the maximum-likelihood criterion. The major difference among these inference methods lies in the message update equation. VB updates messages by complicated digamma functions, which cause bias and slow down message updating (Zeng et al., 2011). GS updates messages by topic labels randomly sampled from the message in the previous iteration. The sampling process does not keep all uncertainty encoded in the previous message. In contrast, BP directly uses the previous message to update the current message without sampling. Similar ideas have also been proposed within the approximate mean-field framework (Asuncion, 2010) as the zero-order approximation of the collapsed VB (CVB0) algorithm (Asuncion et al., 2009). While proper settings of hyperparameters can make the topic modeling performance comparable among different inference methods (Asuncion et al., 2009), we still advocate the BP algorithms because of their ease of use and fast speed. Table 1 compares the message update equations among VB, GS and BP. Compared with BP, VB uses the digamma function Ψ in message update, and GS uses the discrete count of sampled topic labels $n^{-i}_{w,d}$ based on word tokens rather than word index in message update. The Dirichlet hyperparameters α and β can be viewed as the pseudo-messages. The notations $-w$ and $-d$ denote all word indices except w and all document indices except d, and $-i$ denotes all word tokens except the current word token i. More details can be found in our work (Zeng et al., 2011, 2012a).

Because VB and GS have been widely used for learning different LDA-based topic models, it is easy to develop the corresponding BP algorithms for learning these LDA-based topic models by

1. See `http://www.cs.princeton.edu/~blei/lda-c/index.html`.
2. See `http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm`.
3. See `http://nlp.stanford.edu/software/tmt/tmt-0.3/`.
4. See `http://CRAN.R-project.org/package=lda`.
5. See `http://mallet.cs.umass.edu/`.
6. See `http://www.arbylon.net/projects/`.
7. See `http://CRAN.R-project.org/package=topicmodels`.

Inference methods	Message update equations
VB	$\mu_{w,d}(k) \propto \frac{\exp[\Psi(x_{\cdot,d}\mu_{\cdot,d}(k)+\alpha)]}{\exp[\Psi(\sum_k[x_{\cdot,d}\mu_{\cdot,d}(k)+\alpha])]} \times \frac{x_{w,\cdot}\mu_{w,\cdot}(k)+\beta}{\sum_w[x_{w,\cdot}\mu_{w,\cdot}(k)+\beta]}$
GS	$\mu_{w,d,i}(k) \propto \frac{n_{\cdot,d}^{-i}(k)+\alpha}{\sum_k[n_{\cdot,d}^{-i}(k)+\alpha]} \times \frac{n_{w,\cdot}^{-i}(k)+\beta}{\sum_w[n_{w,\cdot}^{-i}(k)+\beta]}$
BP	$\mu_{w,d}(k) \propto \frac{x_{-w,d}\mu_{-w,d}(k)+\alpha}{\sum_k[x_{-w,d}\mu_{-w,d}(k)+\alpha]} \times \frac{x_{w,-d}\mu_{w,-d}(k)+\beta}{\sum_w[x_{w,-d}\mu_{w,-d}(k)+\beta]}$

Table 1: Comparison of message update equations (Zeng et al., 2011).

either removing the digamma function in the VB or without sampling from the posterior probability in the GS algorithm. For example, we show how to develop the corresponding BP algorithms for two typical LDA-based topic models such as ATM and RTM (Zeng et al., 2011).

3. An Example of Using TMBP

TMBP toolbox contains source codes for learning LDA based on VB, GS, and BP (Zeng et al., 2011, 2012a,b,c), learning author-topic models (ATM) (Rosen-Zvi et al., 2004) based on GS and BP, learning relational topic models (RTM) (Chang and Blei, 2010) and labeled LDA (Ramage et al., 2009) using BP. Implementation details can be found in "readme.pdf", which is distributed with the software. Here, we present a demo for the synchronous BP algorithm. After installation, we run `demo1.m` in the Octave/Matlab environment. The results (the training perplexity at every 10 iterations and the top five words in each of ten topics) are printed on the screen:

```
********************
The sBP Algorithm
********************
    Iteration 10 of 500:    1041.620873
    ...
    ...
    Iteration 490 of 500:   741.946849
Elapsed time is 13.246747 seconds.
********************
Top five words in each of ten topics by sBP
********************
design system reasoning case knowledge
model models bayesian data markov
genetic problem search algorithms programming
algorithm learning number function model
learning paper theory knowledge examples
learning control reinforcement paper state
model visual recognition system patterns
research report technical grant university
network neural networks learning input
data decision training algorithm classification
```

Acknowledgments

This work is supported by NSFC (Grant No. 61003154), Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant No. 12KJA520004), the Shanghai Key Laboratory of Intelligent Information Processing, China (Grant No. IIPL-2010-009), and a grant from Baidu.

References

A. Asuncion. Approximate mean field for Dirichlet-based models. In *ICML Workshop on Topic Models*, 2010.

A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *UAI*, pages 27–34, 2009.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

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

J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4(1):124–150, 2010.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proc. Natl. Acad. Sci.*, 101:5228–5235, 2004.

F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Inform. Theory*, 47(2):498–519, 2001.

D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Empirical Methods in Natural Language Processing*, pages 248–256, 2009.

M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *UAI*, pages 487–494, 2004.

J. Zeng, W. K. Cheung, and J. Liu. Learning topic models by belief propagation. *arXiv:1109.3437v4 [cs.LG]*, 2011.

J. Zeng, Z.-Q. Liu, and X.-Q. Cao. A new approch to speeding up topic modeling. *arXiv:1204.0170v1 [cs.LG]*, 2012a.

J. Zeng, Z.-Q. Liu, and X.-Q. Cao. Memory-efficient topic modeling. *arXiv:1206.1147v1 [cs.LG]*, 2012b.

J. Zeng, Z.-Q. Liu, and X.-Q. Cao. Residual belief propagation for topic modeling. *arXiv:1204.6610v1 [cs.LG]*, 2012c.