分散表現を用いたトピック抽出における確率的変分推論法適用

尾崎 花奈　　小林 一郎
お茶の水女子大学 理学部 情報科学科

{ozaki.kana, koba}@is.ocha.ac.jp

1 はじめに

トピックモデルは，文書の中に潜在的に存在するトピックを自動で抽出するためのモデルである．近年最も広く使われているトピック抽出の手法であるLDA(Latent Dirichlet Allocation)[1] は，各文書に潜在トピックがあると仮定し，統計的に共起しやすい単語の集合が生成される要因を，潜在トピックという観測できない確率変数で定式化する．近年 LDA を改良した様々なモデルが提案され，その中でも Das ら [2] によって提案された，単語の分散表現と LDA を組み合わせた Gaussian LDA が注目を集めている．

単語の意味的関係性を事前知識として持つため，従来の LDA に比べ，Gaussian LDA の方が自己相互情報量 (PMI) の値が高くなったと報告している．また，交差検定をした際に訓練データには無く，評価用データに出現する未知語に対してもトピックを推定できるようになった．これにより，従来の LDA においてトピックの推定を諦めていた未知語に対してもトピックの割り当てができるようになった．

Gaussian LDA における事後分布推定方法では，周辺化ギブスサンプリングを用いているが，本稿では SVI (Stochastic Variational Inference)[3][4] を用いることによって，計算時間の大幅な短縮が可能になり，大規模なコーパスに対して効率的な処理が可能になる．

2 関連研究

2.1 Latent Dirichlet Allocation(LDA)

LDA は，Blei ら [1] によって提案されたテキストの生成モデルである．文書中に含まれる潜在トピックに対し，単語が出現する確率分布が定義され，そこから単語が抽出され文書が生成されるという生成モデルに基づき，文書中の潜在トピックを推定する．

LDA の生成モデルをまとめると以下のようになる．

1. for $k = 1$ to K
   (a) Choose topic $\boldsymbol{\beta}_k \sim \mathrm{Dir}(\boldsymbol{\eta})$
2. for each document d in corpus D
   (a) Draw topic distribution $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\alpha})$
   (b) for each word index n from 1 to N_d
      (a) Draw a topic $z_n \sim \mathrm{Categorical}(\boldsymbol{\theta}_{\mathrm{d}})$
      (b) Draw a word $w_n \sim \mathrm{Categorical}(\boldsymbol{\beta}_{\mathrm{z_n}})$

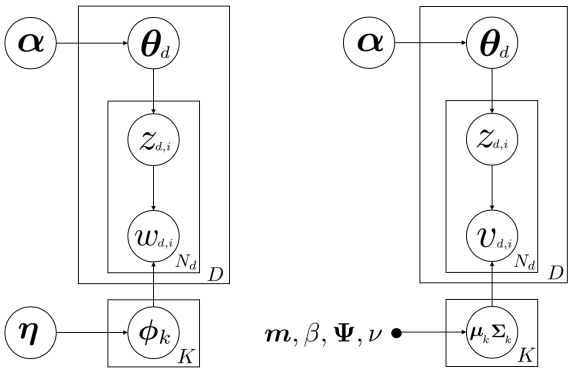
ここで $\boldsymbol{\theta_d}$ は文書 d のトピック分布，$\boldsymbol{\beta}_k$ はトピック k における単語の出現分布を表す．

LDA のグラフィカルモデルを図 1(a) に示す．

2.2 Gaussian LDA

まず，LDA におけるトピックを生成する分布を多次元ガウス分布にするというモデルが Hu ら [5] によって提案された．このモデルに，単語の分散表現を組み合わせたものが Das ら [2] によって提案されたモデルである．Embedding のツールとしては，word2vec [6] を用いている．連続空間に Embedding された単語ベクトルに対し，トピック k を同空間上での多次元ガウス分布としている．これによって，トピックごとの単語分布が連続分布となり，この分布から単語ベクトルが生成される過程がモデル化される．

単語の分散表現を用いることによって，トピック内の意味的結束性が向上し，実験結果として従来の LDA と比較して PMI が上昇することが確認されている．また，トピックごとの単語分布に連続分布を用いることによって，従来の LDA では対応できていなかった未知語に対しても，もう一度モデルでの推定を行うことなしに潜在トピックを割り当てることが可能になっている．



(a) LDA (b) Gaussian LDA

図 1: LDA と Gaussian LDA のグラフィカルモデル

Gaussian LDA の生成モデルをまとめると以下のようになる．

1. for $k = 1$ to K
   - **(a)** Draw topic covariance $\boldsymbol{\Sigma}_k \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}, \mu)$
   - **(b)** Draw topic mean $\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\mu}, \frac{1}{\beta}\boldsymbol{\Sigma}_k)$
2. for each document d in corpus D
   - **(a)** Draw topic distribution $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\alpha})$
   - **(b)** for each word index n from 1 to N_d
     (a) Draw a topic $z_n \sim \mathrm{Categorical}(\boldsymbol{\theta}_{\mathrm{d}})$
     (b) Draw $\boldsymbol{v}_{d,n} \sim \mathcal{N}(\boldsymbol{\mu}_{z_n}, \boldsymbol{\Sigma}_{z_n})$

ここで $\boldsymbol{\theta}_d$ は従来の LDA と同じく文書 d のトピック分布を表すが，$\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ はそれぞれトピック k における多次元ガウス分布の平均と分散を表している．また，$\boldsymbol{v}_{d,n}$ は単語ベクトルを表す．

Gaussian LDA のグラフィカルモデルを図 1(b) に示す．

3 SVI を用いたトピック推定

Gaussian LDA において，事後分布の推定には周辺化ギブスサンンプリングを用いていた．しかし，ギブスサンプリングは実装が簡潔である利点はあるが，計算時間が非常にかかる．

そこで本稿では, 確率的変分近似法 (SVI : Stochastic Variational Inference) [4] を用いることによって，計算時間の大幅な減少を実現し，大規模なデータに対して効率的にトピック解析することを目指す．

変分ベイズにおいては，真の事後分布に対してより簡単な近似分布 $q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ を考え，対数周辺尤度 $p(\boldsymbol{v}|\boldsymbol{\alpha}, \boldsymbol{\zeta})$ の変分下限を最大にする $q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ を求める．

$$\begin{aligned}\log p(\boldsymbol{v}|\boldsymbol{\alpha}, \boldsymbol{\zeta}) &\geq L(\boldsymbol{v}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\zeta}) \\ &\triangleq \mathbb{E}_q[\log p(\boldsymbol{v}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{\alpha}, \boldsymbol{\zeta})] \\ &\quad - \mathbb{E}_q[\log q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})].\end{aligned} \tag{1}$$

平均場近似に基づいて，近似分布 q に対して次のように各確率変数に独立性の仮定をおく．

$$q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = q(\boldsymbol{z})q(\boldsymbol{\theta})q(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{2}$$

単語ごとのトピック割り当て z のパラメータを ϕ, 文書ごとのトピック分布 θ のパラメータを γ，トピックごとの単語分布の平均と分散 $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ のパラメータを $\boldsymbol{\zeta} = (\boldsymbol{m}, \beta, \boldsymbol{\Psi}, \nu)$ とすると，近似分布 q はそれぞれ以下のように表される．

$$\begin{aligned}&q(z_{di} = k) = \phi_{dw_{di}k}; \qquad q(\theta_d) = \mathrm{Dir}(\theta_{\mathrm{d}}|\gamma_{\mathrm{d}}), \\ &q(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = NIW(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k|\boldsymbol{m}_k, \beta_k, \boldsymbol{\Psi}_k, \nu_k).\end{aligned} \tag{3}$$

また，パラメータ $\boldsymbol{\phi}, \boldsymbol{\gamma}$ は以下のように定義される．

$$\begin{aligned}&\phi_{dwk} \propto \exp\{\mathbb{E}_q[\log \theta_{dk}] + \mathbb{E}_q[\log N(\boldsymbol{v}_{dw}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]\}, \\ &\gamma_{dk} = \alpha + \sum_w n_{dw}\phi_{dwk}.\end{aligned} \tag{4}$$

従来の変分近似法における LDA の学習では，文書データ全体に対して繰り返し学習が必要であったが, SVI は文書を逐次的に学習する．$q(\boldsymbol{z}_d), q(\boldsymbol{\theta}_d)$ は各文書ごとに学習される近似事後分布であるので逐次学習を行う必要はなく，逐次学習の対象となるのは $q(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ である．よって，パラメータ $\boldsymbol{\zeta} = (\boldsymbol{m}, \beta, \boldsymbol{\Psi}, \nu)$ の更新において, 確率的自然勾配法を用いた最適化を行う．

n_d 個の単語を含む d 番目の文書において，$\boldsymbol{\zeta}$ は固定して ϕ_d と γ_d の最適化を行う．次に，$\boldsymbol{\zeta}$ の中間パラメータ $\boldsymbol{\zeta}^* = (\boldsymbol{m}^*, \beta^*, \boldsymbol{\Psi}^*, \nu^*)$ を以下の式で求める．

$$\begin{aligned}&{\beta_k}^* = \beta + D\sum_w n_{dw}\phi_{dwk}; \qquad \nu_k^* = \nu + D\sum_w n_{dw}\phi_{dwk}, \\ &\boldsymbol{m}_k^* = \frac{\beta\boldsymbol{m} + D\sum_w n_{dw}\phi_{dwk}\bar{\boldsymbol{v}}_k}{\beta_k^*}, \\ &\boldsymbol{\Psi}_k^* = \boldsymbol{\Psi} + \boldsymbol{C}_k + \frac{\beta D\sum_w n_{dw}\phi_{dwk}}{\beta_k^*}(\bar{\boldsymbol{v}}_k - \boldsymbol{m})(\bar{\boldsymbol{v}}_k - \boldsymbol{m})^{\mathrm{T}}.\end{aligned} \tag{5}$$

ここで,

$$\begin{aligned}&\bar{\boldsymbol{v}}_k = \frac{\sum_w n_{dw}\phi_{dwk}\boldsymbol{v}_{dw}}{\sum_w n_{dw}\phi_{dwk}}, \\ &\boldsymbol{C}_k = D\sum_w n_{dw}\phi_{dwk}(\boldsymbol{v}_{dw} - \bar{\boldsymbol{v}}_k)(\boldsymbol{v}_{dw} - \bar{\boldsymbol{v}}_k)^{\mathrm{T}}.\end{aligned} \tag{6}$$

D はコーパスの数を表しており，$\boldsymbol{\zeta}$ の計算を文書 d の複製 D 個に対して適用することを意味している．この操作によって，パラメータ $\boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\zeta}$ を更新する各イテレーションにおいてコーパス全体を必要とすることがなくなり，大規模なデータに対して逐次的な計算が可能になる．次のイテレーションに用いる $\boldsymbol{\zeta}$ は，$\rho_d \triangleq (\tau_0 + d)^{-\kappa}$，$\kappa \in (0.5, 1]$ で与えられるステップサイズによって，前回のイテレーションの $\boldsymbol{\zeta}$ と更新された $\boldsymbol{\zeta}^*$ に対して重みをかけることによって以下の式で求められる．

$$\boldsymbol{\zeta} = (1 - \rho_d)\boldsymbol{\zeta} + \rho_d \boldsymbol{\zeta}^*. \tag{7}$$

また，q のもとでの $\log \theta_{dk}$ と $\log N(\boldsymbol{v}_{dw}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ の期待値はそれぞれ

$$\begin{aligned}
\mathbb{E}_q[\log \theta_{dk}] &= \Psi(\gamma_{dk}) - \Psi(\sum_{i=1}^{K} \gamma_{di}), \\
\mathbb{E}_q[\log N(\boldsymbol{v}_{dw}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)] &= -\frac{1}{2}\boldsymbol{v}_{dw}^{\mathrm{T}} \langle \boldsymbol{\Sigma}_k^{-1} \rangle \boldsymbol{v}_{dw} \\
&\quad + \boldsymbol{v}_{dw}^{\mathrm{T}} \langle \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k \rangle - \frac{1}{2} \langle \boldsymbol{\mu}_k^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k \rangle \\
&\quad - \frac{1}{2} \langle \log |\boldsymbol{\Sigma}_k| \rangle,
\end{aligned} \tag{8}$$

と表される．ただし，Ψ はディガンマ関数を表し，$\langle\ \rangle$ は期待値を表すものとする．

アルゴリズムは以下のようになる．

Algorithm 1 SVI for Gaussian LDA

Define $\rho_d \triangleq (\tau_0 + d)^{-\kappa}$
Initialize $\boldsymbol{m}, \beta, \boldsymbol{\Psi}, \nu$ randomly.
for $d = 0$ to ∞ **do**
  Estep:
  initialize $\gamma_{dk} = 1$ (The constant 1 is arbitrary.)
  repeat
    Set $\phi_{dwk} \propto \exp\{\mathbb{E}_q[\log \theta_{dk}] + \mathbb{E}_q[\log N(\boldsymbol{v}_{dw}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]\}$
    Set $\gamma_{dk} = \alpha + \sum_w n_{dw}\phi_{dwk}$
  until $\frac{1}{K}\sum_k |change\ in\ \gamma_{dk}| < 0.00001$
  Mstep:
  Compute $\boldsymbol{\zeta}_k^*$ with Eq.(5)
  Set $\boldsymbol{\zeta} = (1 - \rho_d)\boldsymbol{\zeta} + \rho_t \boldsymbol{\zeta}^*$
end for

4 実験

本研究では，単語の分散表現を用いたトピックモデルに SVI を組み込んだモデルを上記のアルゴリズムに従って構築し，文書集合に対してトピック抽出の実験を行った．本稿では，提案手法が従来の LDA に比べて単語の意味的関係性を捉えたトピック抽出をしているのかを評価する実験を行なった．

4.1 実験設定

データセットとして，それぞれ 18846 文書，1740 文書から成る 20Newsgroups[1] と NIPS[2] の 2 つを用いた．ベクトル化された単語のデータとして，Wikipedia で Word2Vec により学習された 50 次元のデータを用い，トピック数 K は 20 から 60 まで 10 ごとに設定した．

文書内のトピック分布の事前分布である Dirichlet 分布のハイパーパラメータ α は $1/K$ とした．また，パラメータの更新式 (7) に用いる τ_0 と κ については，$\tau_0 \in \{1, 4, 16, 64, 256, 1024\}$，$\kappa \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ とした．バッチサイズ S に関しては，文書数の多い 20Newsgroups で $S \in \{4, 16, 64, 256, 1024\}$ とし，文書数の少ない NIPS においては $S \in \{4, 10, 16\}$ とした．提案モデルの実装は Python で行なった．[3]

また，比較対象のモデルとして同じ文書データを用いて従来の LDA を用いて実験をし，Dirichlet 分布のハイパーパラメータ α および η はそれぞれ $1/K$，0.01 とした．

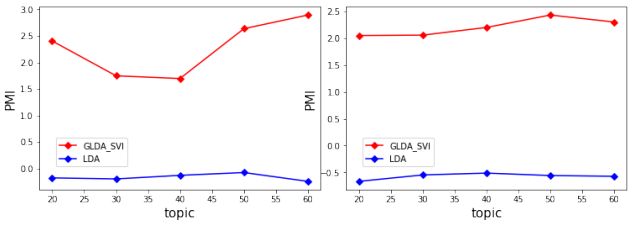
4.2 評価指標

評価指標として，先行研究である Gaussian LDA[2] が用いていた自己相互情報量 (PMI) を用いる．Newman ら [7] によって，トピック内の単語の共起関係を測る指標である PMI がトピックの意味的結束性を表すと提案されている．

コーパスとしては Wikipedia article[4] を用いた．単語 w_i と w_j との共起回数は、同時に同じ文書内に出現した回数とした．トピック k の上位 N 単語の PMI は以下で表される．

$$PMI(k) = \frac{1}{\binom{N}{2}} \sum_{j=2}^{N} \sum_{i=1}^{j-1} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \tag{9}$$

[1] http://qwone.com/~jason/20Newsgroups/
[2] http://www.cs.nyu.edu/~roweis/data.html
[3] https://github.com/KanaOzaki/SVI_GLDA
[4] https://dumps.wikimedia.org/enwiki/latest/



(a) 20Newsgroups (b) NIPS

図 2: 20Newsgroup と NIPS における従来の LDA との各トピック PMI 比較

今回は上位 10 単語の PMI を各トピックの PMI として評価した．

4.3 結果

図 2 は，20Newsgroups と NIPS それぞれについて 20 から 60 の各トピックにおける提案手法と従来の LDA の PMI を表している．提案手法の結果については，様々なハイパーパラメータの組み合わせの中から最も良かったものとして，20Newsgroups は $S = 16$, $\kappa = 1.0$, $\tau_0 = 1024$，NIPS は $S = 4$, $\kappa = 1.0$, $\tau_0 = 1024$ における各トピック上位 10 単語の PMI をプロットした．図 2 からどのトピック数においても従来の LDA よりも PMI の値が上回っていることがわかる．

表 1 は提案手法と従来の LDA で，20Newsgroups におけるトピック数 40 のときのトピックごとの上位 10 単語とその PMI を表している．パラメータは上記と同じ設定であり、PMI が高い順に上位 5 トピックを表示している．表 1 より，20Newsgroups データセットにおいて提案手法で抽出されたトピックの方が全体的に高い PMI を持っていることがわかる．

5 まとめと今後の課題

本研究では，単語の分散表現を取り入れたトピックモデルに おいて，大規模テキストに対応できる効率の良い計算方法を導入する提案を行い，単語の分散表現を取り入れたことによるトピックの意味的結束性の向上を実験により検証した．今後は SVI を用いた提案手法と，Gaussian LDA のモデルにおける事後分布推定にギブスサンプリングを用いたものと比較し、パープレキシティの収束速度を検証することによって提案手法における効率的なトピック推定の検証を行っていく．

表 1: 提案手法と従来の LDA における 20newsgroups データセットに対してのトピック上位 10 単語とその PMI.

提案手法				
cie	geophysics	manning	authenticity	beasts
informatik	astrophysics	neely	veracity	creatures
nos	physics	carney	credence	demons
gn	meteorology	brady	assertions	monsters
nr	astronomy	wilkins	inaccuracies	eleves
sta	geophysical	brett	particulars	spirits
vy	geology	seaver	textual	unicorns
gl	astrophysical	reggie	merits	denizens
cs	chemistry	ryan	substantiate	magical
ger	microbiology	wade	refute	gods
6.6429	6.2844	5.3070	5.0646	4.3270
Multinomial LDA topics				
drive	ax	subject	data	south
disease	max	lines	doctors	book
hard	a86	server	teams	lds
scsi	0d	organization	block	published
drives	1t	spacecraft	system	adl
disk	giz	spencer	spave	armenian
subject	3t	program	output	books
daughter	cx	space	pool	documents
unit	bh	software	resources	subject
organization	kt	graphic	bits	information
2.3514	2.2500	1.3700	1.1216	1.0528

参考文献

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. (2003) Latent dirichlet allocation, J. Mach. Learn. Res., 3:993-1022, March.

[2] Rajarshi Das, Manzil Zaheer, and Chris Dyer. (2015) Gaussian LDA for topic models with word embeddings. In Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics.

[3] Matthew D. Hoffman, David M. Blei, Francis Bach. (2013) Online Learning for Latent Dirichlet Allocation. In Advances in Neural Information Processing Systems.

[4] Matthew D. Hoffman, David M. Blei, Chong Wang, John Paisley. (2013) Stochastic variational inference. The Journal of Machine Learning Reseach, 14(1):1303-1347.

[5] Pengfei Hu, Wenju Liu, Wei Jiang, and Zhanlei Yang. (2012) Latent topic model based on Gaussian-LDA for audio retrieval. In Pattern Recognition, volume 321 of CCIS, pages 556-563. Springer.

[6] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. (2013) Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 746-751, Atlanta, Georgia, June.

[7] David Newman, Sarvnaz Karimi, and Lawrence Cavedon.(2009) External evaluation of topic models. pages 11-18, December.