The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

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への取り組み

A Study on Stochastic Variational Inference for Topic Modeling with Word Embeddings

尾崎花奈 [*1] 小林 一郎 [*2]
Kana Ozaki Ichiro Kobayashi

[*1]お茶の水女子大学理学部情報科学科 [*2]お茶の水女子大学基幹研究院自然科学系
Ochanomizu University Ochanomizu University

Probabilistic topic models based on latent Dirichlet Allocation is widely used to extract latent topics from document collections. In recent years, a number of extended topic models have been proposed, especially Gaussain LDA(G-LDA) has attracted a lot of attention. G-LDA integrates topic modeling with word embeddings by replacing discrete topic distribution over word types with multivariate Gaussian distribution on the word embedding space. This can reflect semantic information into topics. In this paper, we use a G-LDA for our base topic model and apply Stochastic Variational Inference (SVI), an efficient inference algorithm, to estimate topics. This encourages the model to analyze massive document collections, including those arriving in a stream.

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．代表的な手法である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．Das ら [2] によって提案された Gaussian LDA は，LDA の改良モデルであり、LDA に単語の分散表現 (Word embedding) を組み合わせている．従来の LDA ではトピックごとの単語分布をカテゴリ分布としていたのに対し，Gaussian LDA では embedding 空間上の多次元ガウス分布としている．単語の意味的関係性を事前知識として持つため，Gaussian LDA の方がトピックごとの自己相互情報量 (PMI) の値が高くなったと報告している．

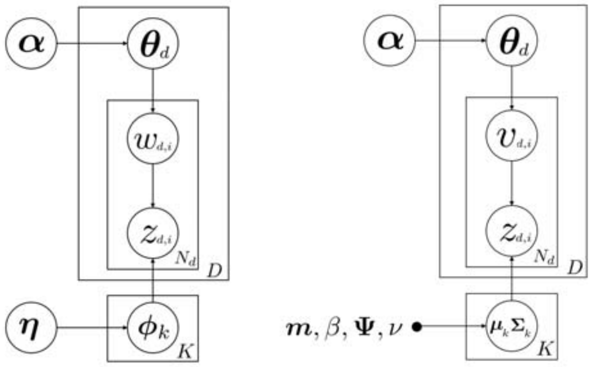
Gaussian LDA における事後分布推定方法では，周辺化ギブスサンプリングを用いているが，本稿では SVI (Stochastic Variational Inference)[3][4] を用いることによって，計算時間の大幅な短縮を行い，大規模なコーパスに対しての効率的な処理を目指す．

2. Gaussain 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 Gaussian LDA [2] 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．

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，トピックの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．

Gaussian LDA の生成モデルは以下のようになる．



(a) LDA (b) Gaussian LDA

図 1: LDA と Gaussian LDA のグラフィカルモデル

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．

LDA のグラフィカルモデルを図 1(a) に，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．

連絡先: 尾崎花奈，お茶の水女子大学大学 情報科学科，〒 112-8610 東京都文京区大塚 2-1-1, g1420513@is.ocha.ac.jp

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．

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

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．

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

アルゴリズムは以下のようになる．

Algorithm 1 SVI for Gaussian LDA

```
Define ρ_d ≜ (τ_0 + d)^{-κ}
Initialize m, β, Ψ, ν randomly.
for d = 0 to ∞ do
  Estep:
  initialize γ_dk = 1 (The constant 1 is arbitrary.)
  repeat
    Set φ_dwk ∝ exp{E_q[log θ_dk] +
                 E_q[log N(v_dw|μ_k, Σ_k)]}
    Set γ_dk = α + Σ_w n_dw φ_dwk
  until (1/K) Σ_k |change in γ_dk| < 0.00001
  Mstep:
  Compute ζ*_k with Eq.(5)
  Set ζ = (1 − ρ_d)ζ + ρ_t ζ*
end for
```

4. 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なった．今後は，実験を通じて提案手法の正当性を検証していく．

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. (2010) 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.