確率的トピックモデルを用いた医学生物学文献情報に基づく仮説生成

麻生 竜矢† 江口 浩二†
† 神戸大学大学院工学研究科情報知能学専攻

1 はじめに

近年，医学生物学などの学術分野において，大量の電子文献に蓄積された知見から潜在的な仮説を生成する技術への要求が高まっている．この目的の下で，生物学的知識，特にタンパク質間相互関係に関する知識の抽出のため，統計的手法に基づいたアプローチとして確率的トピックモデルを適用し，分類精度とランキング精度の観点から，その有用性を示す．特に潜在的ディリクレ配分法（Latant Dirichlet Allocation: LDA）による確率的トピックモデルは，上述のようなタスクに関する有効性という観点からはこれまで検討されてこなかった．本稿ではLDAの推定手法としてCollapsed変分ベイズ法を適用し，一般的なCollapsed Gibbs Sampling法によるLDAと確率的潜在意味インデクシング法（probabilistic Latant Semantic Indexing: pLSI）との比較を行う．

本稿はTREC Genomic Trackの一部を用いた実験の結果を報告するものである．

2 潜在的ディリクレ配分法

本節では文献[2]で提案された，Collapsed変分ベイズ法によるLDAモデル[1]について述べる．

LDAモデル LDAモデルによる文書生成過程[1]は以下の手順に従う．

(1) ハイパーパラメータ α のディリクレ分布から各文書 j について θ_j をサンプリング

(2) ハイパーパラメータ β を与えたディリクレ分布から各トピック k について ϕ_k をサンプリング

(3) 文書 j 内の N_j 個の語 x_i それぞれに対して
(a) パラメータ θ_j を与えた多項分布からトピック z_i をサンプリング
(b) パラメータ ϕ_{z_i} を与えた多項分布から語 x_i をサンプリング

また，LDAの全パラメータと確率変数の同時分布は，次のようになる．

$$p(\mathbf{x},\mathbf{z},\theta,\phi|\alpha,\beta) = \prod_{j=1}^{D}\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}\prod_{k=1}^{K}\theta_{jk}^{\alpha-1+n_{jk\cdot}} \times \prod_{k=1}^{K}\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\prod_{w=1}^{W}\phi_{kw}^{\beta-1+n_{\cdot kw}} \quad (1)$$

このとき，K はトピック数，D は文書数，W は総語彙数を示す．また，n_{jkw} は文書 j 中に存在するトピック k に属する単語 w の数であり，ドット（$\cdot$）は一致するインデックスが総計されることを意味する．つまり $n_{\cdot kw}=\sum_j n_{jkw}$，$n_{jk\cdot}=\sum_w n_{jkw}$ である．

実際に単語 $\mathbf{x}=\{x_{ji}\}$ が与えられたとき，潜在的トピックインデックス $\mathbf{z}=\{z_{ij}\}$，混合比 $\theta=\{\theta_j\}$ とトピックパラメータ $\phi=\{\phi_k\}$ 上の事後分布を計算するようなベイズ推定を考える．その実現方法がCollapsed Gibbs Sampling法であり，また以下で解説するCollapsed変分ベイズ法である．

Collapsed変分ベイズ法 Collapsed変分ベイズ（CVB）推定とは，パラメータ ϕ，θ を周辺化することで推定精度を高めたアルゴリズムである．このアルゴリズムはCollapsed Gibbs Sampling法とは異なり，潜在変数でパラメータの依存性をモデル化している．

Collapsed変分ベイズ法において，文書 j の i 番目の単語におけるトピック k の分布のパラメータ γ_{ijk} は以下の式に基づいて逐次的に更新される．

$$\gamma_{ijk} \propto (\alpha+E_q[n_{jk\cdot}^{\neg ij}])(\beta+E_q[n_{\cdot kx_{ij}}^{\neg ij}])(W\beta+E_q[n_{\cdot k\cdot}^{\neg ij}])^{-1} \exp\left(-\frac{Var_q(n_{jk\cdot}^{\neg ij})}{2(\alpha+E_q[n_{jk\cdot}^{\neg ij}])^2}-\frac{Var_q(n_{\cdot kx_{ij}}^{\neg ij})}{2(\beta+E_q[n_{\cdot kx_{ij}}^{\neg ij}])^2}+\frac{Var_q(n_{\cdot k\cdot}^{\neg ij})}{2(W\beta+E_q[n_{\cdot k\cdot}^{\neg ij}])^2}\right) \quad (2)$$

このとき添字 $\neg ij$ が付いたものは x_{ij} や z_{ij} を除いた変数や頻度に相当する．

タンパク質間の類似度 本節では文献[3]で提案されている類似度計算手法に対して，より有用であると考えられる計算手法を提示している．

LDAなどのトピックモデルを利用すれば，あるタンパク質のペアが将来において，文書中に現れる尤度を計算することは，たとえそのペアがそれまでどの文書にも存在しなかったとしても，可能である．ただし，個々のタンパク質は訓練データのいずれかの文書において既に出現しているものとする．

LDAに基づいた，2つのタンパク質間の類似度[3]は，

$$Sim1(e_i,e_j)=p(e_i|e_j)/2+p(e_j|e_i)/2 \quad (3)$$

を用いることで測定が可能であるが，この方法の場合，類似度が少数の頻出するタンパク質の側に依存してしまう，という問題が考えられる．例えば，片方の条件付き確率が極端に大きな値になったとき，もう片方の確率がほとんど0であった場合でも，二つの値の平均をとる計算手法では，導出される結果が充分大きな値になってしまう．そこで本論文では，より正確な類似度を導出するため，式（3）を以下のように改良する．

$$Sim2(e_i,e_j)=p(e_i|e_j)\times p(e_j|e_i) \quad (4)$$

二つの値の積をとるこの方法であれば，上記のような場合でも計算結果は0に近い値になる．

3 実験

文献データセット データセットとして2004年から2005年のTREC Genomics Track[3]で使用された

Hypothesis Generation Based on Biomedical Literature Using Probabilistic Topic Models
Tatsuya ASOU† and Koji EGUCHI†, †Department of Computer Science and Systems Engineering, Kobe University

[3] http://ir.ohsu.edu/genomics/

TREC コレクションの一部を利用した．この内，出版年が 2002 年である文献（33,000 件）のタイトルと要旨を訓練データ，2003 年の文献（33,000 件）のものをテストデータとしている．

文献データの前処理　訓練データに対してモデル推定を行う前に，両方のデータにおいて以下に述べる幾つかの処理を行った．まず 419 種類のストップワード（“this” や “a” など，推定において参考にならない単語）を除去した．また，訓練データにおいて 10 件未満の文書にしか出現していない単語を除去した．また，TREC コレクションではタンパク質にタグ付けなどされていないので，GENIA タガー[4] と呼ばれる解析ツールを使用してタンパク質とその他の語の分類を自動で行った．なお，このツールによるタンパク質へのタグ付けの精度は 70 %程度である．

評価データの作成　分類精度とランキング精度の評価用データとして，タンパク質ペアのセットを二種類用意した．1 つ目のセット「正解ペア」データは，同一文献中での共起が訓練データでは確認されなかったが，テストデータでは確認されたタンパク質ペアのデータセットである．ただし，訓練データに出現しないタンパク質名を含んだペアは除外した．もう 1 つのセット「不正解ペア」データとは，訓練データとテストデータの双方で一度も共起が確認されていないエンティティ対のデータセットである．

実験の流れ　上記の前処理を行った後，Collapsed 変分ベイズ法による LDA モデルと Collapsed Gibbs Sampling 法による LDA モデル，そして pLSI をそれぞれ用いてモデル推定を行った．それから生成された三種類の推定モデルを用いて，二種類の計算手法で全タンパク質ペアの類似度をそれぞれ計算した．最後に分類精度とランキング精度で評価した．ただし，評価値については初期値をランダムに設定して 50 回のモデル推定を行ったときの，それぞれの精度の期待値を最終的な値とする．ただし，pLSI については結果にばらつきが生じないので 1 回分の分類精度を示している．

表 1: 分類精度

	CVB-LDA	CGS-LDA	pLSI
K=10	0.6310	**0.6318**	**0.6075**
K=50	**0.6434**	0.5359	0.5829
K=100	0.6383	0.5669	0.5648
K=300	0.6317	0.5504	0.5293

表 2: ランキング精度

	CVB-LDA	CGS-LDA	pLSI
K=10	0.6651	**0.6745**	**0.6347**
K=50	0.6895	0.6574	0.5977
K=100	**0.6905**	0.6443	0.5606
K=300	0.6890	0.6262	0.5308

[4] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=GENIA+Tagger

提案手法の適用と評価結果　分類精度（Classification Accuracy: CA）とランキング精度（Average Precision: AP）の両方において，Collapsed 変分ベイズ法によって推定された LDA によって得られた改善は，Collapsed Gibbs Sampling 法で推定された LDA の場合と比較して統計的に有意であることが確認できた．このとき，実験において最良の評価値であったトピック数を条件として両手法を比較した．つまり，分類精度の比較では，$K = 50$ のときの Collapsed 変分ベイズ法を $K = 10$ のときの Collapsed Gibbs Sampling 法に対して比較し，ランキング精度の比較では，$K = 100$ のときの Collapsed 変分ベイズ法を $K = 10$ のときの Collapsed Gibbs Sampling 法に対して比較した．また，検定には Wilcoxon 符号付順位検定（両側）を用いて，有意水準は 5%とした．また，Collapsed 変分ベイズ LDA（CVB-LDA）と Collapsed Gibbs Sampling LDA（CGS-LDA）は，分類精度とランキング精度のいずれの観点からも pLSI より高精度であることがわかった．

表 3: 類似度計算手法の比較結果（CVB-LDA）

		K=10	K=50	K=100	K=300
CA	*Sim*1	0.6310	0.6434	0.6383	0.6317
	*Sim*2	0.6351*	0.6467*	0.6398*	0.6305*
AP	*Sim*1	0.6651	0.6895	0.6905	0.6890
	*Sim*2	0.6719*	0.6947*	0.6940*	0.6892

*Sim*2 に関する *Sim*1 を基準とした改善について有意差が認められた場合に * を付した．

また，二種類の LDA モデルを用いて（3）式と（4）式の二種類の計算手法によって各エンティティ対の類似度を計算し，それぞれの結果に対して分類精度とランキング精度を測定した．その結果，全体的に（4）式を用いた方が良好な結果が得られた．こちらも，統計的には概ね有意であることを確認している．その一例として，Collapsed 変分ベイズ LDA で比較した際の結果を表 3 に示す．

本稿では文献データだけを利用したが，生物学的実験に基づくタンパク質相互作用ネットワークと，文献から得たタンパク質相互関係ネットワークを組み合わせることも現在検討中である．

謝辞　本研究の一部は，科学研究費補助金基盤研究 (B) (20300038) の援助による．

参考文献

[1]Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent dirichlet allocation, *Journal of Machine Learning Research*, Vol.3, pp. 993–1022(2003).

[2]Teh, Y.W., Newman, D. and Welling, M.: A collapsed variational bayesian inference algorithm for latent dirichlet allocation, *Advances in Neural Information Processing Systems* Vol.16 (2007).

[3]Newman, D., Chemudugunta, C., Smyth, P. and Steyvers, M.: Statistical entity-topic models, *Proceeding of the 12th ACM SIGKDD international conference on Knowledge discoverry and data mining*, Philadelphia, Pennsylvania, pp. 680–686 (2006).