Vol.2019-BIO-60 No.1
2019/12/2

対象者の属性情報を考慮した, 細菌組成データに対するノンパラメトリックベイズトピックモデル

奥井　佑[1,a)]

概要：本研究では, 対象者の属性を考慮可能な潜在ディリクレ配分モデル (LDA) として提案されているディリクレ多項回帰モデルを伴う LDA (DMR トピックモデル) を, 棒折過程を用いてノンパラメトリックベイズに拡張したモデルを提案し, 実際の細菌データをもとに性能を検証した. 結果, 提案法では既存手法と比べて属性とより関連したトピックが抽出可能であるとともに, データから自動的にトピック数を推定可能となることが示唆された.

A Bayesian Nonparametric Topic Model for Microbiome Data Using Subject Attributes

1. はじめに

次世代シークエンサーを用いた細菌データに対する 16Sr-RNA 遺伝子解析の普及により, 細菌データと臨床的なアウトカムとの関連を分析する研究が多く行われるようになった. 潜在ディリクレ配分モデル (LDA)[1] も分析手法の一つとして, 細菌データから細菌コミュニティを抽出する方法として近年用いられ始めている. 細菌コミュニティを抽出する際, 性年齢等の各対象者の属性を考慮可能であればモデルとしてより望ましいと考えられ, 対象者の属性を考慮可能な LDA としてディリクレ多項回帰モデルを伴う LDA (DMR トピックモデル) が提案されている. また, LDA のトピック数をデータから自動的に推定する方法としてノンパラメトリックベイズモデルが一般的に用いられ, DMR トピックモデルをノンパラメトリックベイズモデルに拡張したモデル:Hierarchical Dirichlet Scaling Process (HDSP)[2] もすでに提案されている. 一方で, HDSP は正規ガンマ過程をもとにトピック分布を生成するが, データから自動的にトピック数を推定可能であるかは不明である. 本研究では, 新たに棒折過程を用いてトピック割合を生成するノンパラメトリックベイズ DMR トピックモデルを提案し, 実際の細菌データをもとに性能を検証する.

[1] 九州大学病院メディカル・インフォメーションセンター
Maidashi 3-1-1, Higashi-ku, Fukuoka 812-8582, Japan
[a)] task1000@gmail.com

2. 提案法

提案法のデータ生成過程を示す.[3]

(1) 各トピック k について,
- (a) $\boldsymbol{\lambda}_k \sim Normal(0, \sigma^2 \boldsymbol{I})$
- (b) $v_k \sim Beta(1, \gamma)$
- (c) $\beta_k = v_k \prod_{i=1}^{k-1}(1 - v_i)$
- (d) $\boldsymbol{\phi}_k \sim Dirichlet(\boldsymbol{\eta})$

(2) 各対象者 d について,
- (a) α_{dk}=expit $(\boldsymbol{\lambda}_k{}' \boldsymbol{x}_d)$.
- (b) $\pi'_{dk} \sim Beta(\xi\beta_k, \xi(1 - \sum_{l=1}^{k}\beta_k))$
- (c) $\pi_{dk} = \pi'_{dk}\alpha_{dk}\prod_{l=1}^{k-1}(1 - \pi'_{dl}\alpha_{dl})$
- (d) 対象者 d の n 番目のシークエンスリードについて,
  - (i) $z_{dn} \sim Multinomial(1, \boldsymbol{\pi}_d)$
  - (ii) $w_{dn} \sim Multinomial(1, \boldsymbol{\phi}_{z_{dn}})$

ここで, z_{dn} は対象者 $d(d=1,\ldots,D)$ の $n(n=1,\ldots,N_d)$ 番目のシークエンスリードのトピックを表し, w_{dn} は対象者 d の n 番目のシークエンスリードの菌種を表すとする. π_{dk} は対象者 d のトピック $k(k=1,\ldots,K)$ におけるトピック割合を表す. $\boldsymbol{\phi}_k$ はトピック k 上の細菌分布 (文書データに対する単語分布に相当) であり, $\boldsymbol{\eta}$ は $\boldsymbol{\phi}_k$ のハイパーパラメータであるとする. β_k は階層ディリクレ過程の 1 段階目の棒折過程により生成される各トピック k のトピック割合のハイパーパラメータを表す. γ は β_k のハイパーパラメータであり, ξ は 2 段階目の棒折過程におけるハイパーパラメー

表 1: いずれかの対象者においてトピック割合が上位となったトピックの数, およびそれらトピックのトピック割合の平均

			事前に設定するトピック数の上限				
データ	トピック	手法	10	20	30	40	50
”妊娠”データ	最も割合が大きいトピック	LDA	10(0.56)	20(0.47)	25(0.43)	35(0.44)	42(0.48)
		HDSP	10(0.77)	16(0.72)	15(0.72)	16(0.72)	18(0.71)
		Proposed model	3(0.82)	5(0.63)	3(0.66)	4(0.66)	4(0.70)
	上位 2 つのトピック	LDA	10(0.77)	20(0.69)	30(0.62)	39(0.64)	50(0.67)
		HDSP	10(0.91)	19(0.87)	24(0.85)	29(0.86)	32(0.85)
		Proposed model	8(0.97)	6(0.89)	6(0.90)	6(0.91)	9(0.89)
”喫煙”データ	最も割合が大きいトピック	LDA	8(0.86)	16(0.80)	24(0.79)	22(0.74)	20(0.76)
		HDSP	9(0.29)	16(0.23)	15(0.23)	17(0.23)	19(0.21)
		Proposed model	6(0.82)	7(0.77)	9(0.76)	7(0.79)	7(0.43)
	上位 2 つのトピック	LDA	10(0.97)	19(0.94)	29(0.93)	32(0.93)	40(0.91)
		HDSP	10(0.49)	19(0.38)	22(0.37)	28(0.37)	26(0.35)
		Proposed model	10(0.95)	12(0.92)	14(0.91)	14(0.93)	9(0.63)

タを表す.$\boldsymbol{x}_d$ は対象者 d の共変量ベクトルであり, $\boldsymbol{\lambda}_k$ は各トピック k の回帰係数を表す. π'_{dk} は 2 段階目の棒折過程により生成される割合であり, 提案法では各対象者の属性情報を含むパラメータ α_{dk} と掛け合わすことでトピック割合を生成する.

属性情報を棒折過程に組み込む方法はカーネル棒折過程とも呼ばれ, この方法を使うことで各対象者のトピック割合の大きさは属性の大きさに直接的に影響されることになる. また, 棒折過程をトピック分布の生成に用いることでトピック割合が少数のトピックに集中するようになることが期待できる. トピック割合の大きさのばらつきはハイパーパラメータ ξ により決定されるが, ξ をデータから推定する場合, トピック割合が大きいトピックはデータからの必要性に応じて制限されることが期待できる.

3. 性能評価

実際の細菌データをもとに提案法の性能を評価した. 評価するモデルとして, 一般的な LDA と HDSP, 提案法を用いた. 性能評価の項目として, 得られたトピックで対象者の属性をどの程度予測できるかを AUC をもとに評価した. また, データからトピック数を自動的に推定できているかを評価するため, いずれかの対象者において一番, あるいは二番目にトピック割合が大きいトピックの数, およびそれらトピックが各対象者のトピック割合に占める割合の平均を集計した. もし, 各手法がデータから自動的にトピック数を推定可能であるならば, トピック割合が大きい上位トピックの数や割合は事前に設定するトピック数によらず一定になると考えられる. 事前に設定するトピック数の上限は 10, 20, 30, 40, 50 と変化させ実施した.

2 種類のデータを用いて提案法の性能を評価した. 1 つ目のデータは妊娠者と非妊娠者の膣内細菌の環境の違いを調べた研究のデータで”妊娠”データとする. データは R パッケージ:NBZIMM[4] から入手した. 属性情報として Nugent スコアと, 妊娠有無の情報を用いた. もう一つのデータは喫煙者と非喫煙者の呼吸器官の細菌環境の違いを調べた研究のデータで”喫煙”データと表記する. R パッケージ:Gunifrac[5] から入手した. 属性情報として, 性別と喫煙有無を用いた.

表 1 が各対象者において一番, あるいは二番目にトピック割合が大きいトピックの数, およびそれらトピックが各対象者のトピック割合に占める割合の平均の結果である. 多くの場合, トピック割合の平均の値は高く, これら上位のトピックが各対象者のトピック内で多くの割合を占めていることがわかる. そのうえで, 提案法においては, 上位トピックの数は事前に設定するトピック数の上限によらず比較的一定であり, その数も LDA や HDSP と比べて小さかった. LDA や HDSP については, 事前に設定するトピックの数を大きくすると上位トピックの数も上昇する傾向がみられた. AUC についても, 提案法の値が他の手法よりも大きくなるケースが多かった.

参考文献

[1] Blei, D., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation, *Journal of Machine Learning Research*, Vol.3, pp.993-1022 (2003).
[2] Kim D., Oh A: Hierarchical Dirichlet scaling process, *Machine Learning*, Vol.106,No.3 pp.387-418 (2017).
[3] Okui T: A Bayesian Nonparametric Topic Model for Microbiome Data Using Subject Attributes, *IPSJ Transactions on Bioinformatics*, (Accepted) .
[4] NBZIMM (online),
available from ⟨https://github.com/nyiuab/NBZIMM⟩ (accessed 2019-7-31)
[5] Chen, J.: package ‘GUniFrac’ Generalized UniFrac distances for comparing microbial communities. Permutational multivariate analysis of variance using multiple distance matrices (online),
available from ⟨https://cran.r-project.org/web/packages/GUniFrac/index.html⟩