修士論文 要旨

生成モデルベースの隠れマルコフモデルと生体データへの応用

数理情報学専攻 48226218　谷本 佑馬
指導教員　小林 徹也 教授

1　はじめに

システムの隠れた状態の推定は普遍的な問題である. 生物学や細胞生理学においては, 生命システムの複雑な状態や運動を理解するために隠れ状態推定は重要である. 例えば, 推測された隠れ状態から生命システムの正常/異常な状態を区別することができる. その応用は表現型スクリーニングや薬剤耐性の同定など多岐にわたる. しかし, これまでの生命システムに対する隠れ状態推定の試みは, 遺伝子の蛍光イメージングや細胞の分裂周期, 分裂長 [8] といった単純な観測データに限られている. 近年, イメージング技術の発展により, 細胞組織の 2 次元・3 次元動画データを取得することが可能になっている. 隠れ状態推定と深層学習アルゴリズムを組み合わせることで, 動画データやラマンスペクトル [7, 9] といったより複雑な高次元データから隠れ状態の抽出を行うことができると考えられる.

2　既存手法

隠れマルコフモデル (hidden Markov model; HMM) [1] は, 隠れ状態推定を目的とした確率モデルである. HMM を用いることで, 観測データから背景にあるシステムの状態やその確率的なダイナミクスの推測を行うとともに, 観測が生成されるメカニズムを解明することができる. 特に, 隠れ状態から生成される観測分布を複数の正規分布の線形和とした, ガウス混合モデル (Gaussian mixture model; GMM) ベースの HMM (GMM-HMM) が広く知られており, 音声認識を中心に幅広い分野で応用されてきた.

長さ T の D 次元観測系列 $\boldsymbol{x} = \{\boldsymbol{x}_0, \dots, \boldsymbol{x}_{T-1}\} \subset \mathbb{R}^D$ と状態系列 $\boldsymbol{s} = \{s_0, \dots, s_{T-1}\}$ を考える. 隠れ状態 $s_t \in \{0, \dots, N-1\}$ は直接観測できない潜在的な確率変数である. 状態系列はマルコフ連鎖であり, 状態遷移確率 $A_{ij} = p(s_t = j \,|\, s_{t-1} = i)$ と初期分布 $\pi_i = p(s_0 = i)$ によって決定づけられる. また状態 i における観測分布を $b_i(\boldsymbol{x}_t) = p(\boldsymbol{x}_t \,|\, s_t = i)$ とおく. GMM-HMM では観測分布として GMM が用いられる:

$$b_i(\boldsymbol{x}_t) = \sum_{k=0}^{M-1} w_{ik}\, \mathcal{N}(\boldsymbol{x}_t \,|\, \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}). \tag{1}$$

ここで, w_{ik} は混合係数, $\boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}$ はそれぞれ状態 i に付随する k 番目の正規分布の平均ベクトル, 分散共分散行列である.

近年, 機械学習の発展に伴い, HMM と機械学習を組み合わせたモデルが提案されている. フローベース生成モデルである Normalizing flow は確率分布推定の文脈で提案された [3]. Normalizing flow は, 正規分布などの単純な分布に可逆かつ微分可能な変換を繰り返し施すことで, より複雑な分布を生成するモデルである. $\boldsymbol{Z} \in \mathbb{R}^D$ は単純な確率分布 (本研究では正規分布) $p_{\boldsymbol{Z}} : \mathbb{R}^D \to \mathbb{R}$ に従う確率変数とする. また $\boldsymbol{g} : \mathbb{R}^D \to \mathbb{R}^D$ を可逆な関数とし, $\boldsymbol{X} = \boldsymbol{g}(\boldsymbol{Z})$ の関係が成立するとする. このとき, 変数変換の公式から $\boldsymbol{X}$ の確率密度関数は次のように書ける:

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = p_{\boldsymbol{Z}}(\boldsymbol{f}(\boldsymbol{x}))|\det \mathbf{D}\boldsymbol{f}(\boldsymbol{x})|, \tag{2}$$

ここで $\boldsymbol{f}$ は $\boldsymbol{g}$ の逆関数, $\mathbf{D}\boldsymbol{f}(\boldsymbol{x})$ は $\boldsymbol{f}$ のヤコビ行列である. 関数 $\boldsymbol{g}$ は単純な分布 $p_{\boldsymbol{Z}}$ からより複雑な分布 $p_{\boldsymbol{X}}$ へ (”生成方向”) の変換である. また逆関数 $\boldsymbol{f}$ はデータの複雑な分布から基本的な分布へ (”正規化方向”) の変換である.

Normalizing flow 混合モデル (Normalizing flow mixture model; NMM) [10] は複数の Normalizing flow 分布の線形和として定義される. この NMM をベースとした HMM (NMM-HMM) [4, 5, 11] は, 各状態における観測分布を NMM によって推定するモデルである:

$$b_i(\boldsymbol{x}_t) = \sum_{k=0}^{M-1} w_{ik}\, p_{\boldsymbol{Z}}(\boldsymbol{f}_{ik}(\boldsymbol{x}_t))|\det \mathbf{D}\boldsymbol{f}_{ik}(\boldsymbol{x}_t)|, \tag{3}$$

ここで, $\boldsymbol{f}_{ik}$ は状態 i の NMM に付随する k 番目の Normalizing flow である. NMM-HMM は従来の GMM-HMM と比較し表現力が高く, データの複雑な多様体を記述できると期待される. 一方, MNIST などの画像データセットに対する教師なし分類において, NMM は GMM よりも精度が低くなることが報告されている [2]. これは, NMM はモデルの複雑度が大きく, パラメータの学習において局所的な最適解に陥りやすいからであると考えられる. NMM をベースとする NMM-HMM でも同様に精度が低くなってしまう可能性がある. そこで本研究では NMM-HMM の安定な学習を可能にする学習アルゴリズムの提案を行う.

3 本研究で提案するアルゴリズム

本節では, 提案手法である GMM-NMM-HMM の学習アルゴリズムについて説明する. このモデルは, 学習した GMM-HMM をベースとして NMM-HMM の学習を行うことで, NMM-HMM の初期パラメータを最適解に近づけ学習の安定化を図るとともに, ベースの GMM-HMM からの精度向上を目指す. なお, 類似モデルとして XGB-HMM [12] が挙げられる.

$\gamma_t(i,k)$ は Baum-Welch アルゴリズムによって計算される, 状態 i の GMM に付随する k 番目の正規分布から観測 $\boldsymbol{x}_t$ が生成される確率である.

$$X_{ik} = \{\boldsymbol{x}_t \,|\, \gamma_t(i,k) > p_{\text{threshold}}\}, \tag{4}$$

とおくと, これは状態 i の GMM に付随する k 番目の正規分布から生成された可能性が高い観測の集合である. この観測集合 X_{ik} を用いて, NMM-HMM の flow $\boldsymbol{f}_{ik}$ の学習を行う. $\boldsymbol{f}_{ik}$ のパラメータを $\boldsymbol{\Phi}_{ik}$ とし, 目的関数を

$$\begin{aligned} Q_{ik}(\boldsymbol{\Phi}_{ik}) &= \sum_{\boldsymbol{x}_t \in X_{ik}} p_{\boldsymbol{X}}(\boldsymbol{x}_t \,|\, s_t = i, m_{ti} = k; \boldsymbol{\Phi}_{ik}) \\ &= \sum_{\boldsymbol{x}_t \in X_{ik}} p_{\boldsymbol{Z}}(\boldsymbol{f}_{ik}(\boldsymbol{x}_t)) |\det \mathbf{D}\boldsymbol{f}_{ik}(\boldsymbol{x}_t)|, \end{aligned} \tag{5}$$

と定める. 各 (i,k) に対して Q_{ik} の最大化を行うことで, NMM-HMM の flow を個別に学習することができる. 以上が GMM-HMM に基づいた NMM-HMM の事前学習の流れである. GMM-NMM-HMM の学習アルゴリズム全体を以下に示す:

(i) GMM-HMM で学習を行い, パラメータ $\boldsymbol{\lambda} = \{\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ を初期化する.
(ii) Baum-Welch アルゴリズムを用いて $\gamma_t(i,k)$ を計算する.
(iii) 各 (i,k) に対して式 (5) の $Q_{ik}(\boldsymbol{\Phi}_{ik})$ を最大化する.
(iv) NMM-HMM の学習を行う.

4 結果

既存手法と提案手法の精度の比較するために, 手書き文字のデータセットである MNIST と EMNIST を用いた検証を行った. NMM-HMM と提案手法の比較では, 提案手法の方が正解率は高く, 特に潜在変数空間の次元が高い場合, その標準偏差は小さくなった. このことから, 提案手法は高い精度で, かつ初期値による学習のばらつきを抑え安定的に NMM-HMM を学習できる有用なモデルであることが示唆される. また GMM-HMM と提案手法の比較では, 特に潜在変数空間の次元が高い場合, 提案手法は GMM-HMM からの一定の精度向上が見られた. ただし, 精度向上には限度があり, 元の GMM-HMM の局所解から抜け出せているわけではないと考えられる.

生体データへの応用として, 動画データからの運動状態推定を行った. *Dictyostelium* に対しては, 細胞の特徴的な形状に対応して隠れ状態が分類された. さらに, HMM による隠れ状態の遷移として, 細胞が突起を形成しながら運動する動態を捉えられることが示された. 一方, *C. elegans* に対しては, 方向転換を行う様子や移動時の体の曲線の位相変化を, 隠れ状態間の状態遷移として表現できることを示した.

表 1. MNIST データに対する予測精度. 10 回分の正解率の平均と標準偏差を示す.

	潜在空間の次元		
	4	8	16
GMM-HMM	0.85 ± 0.06	0.85 ± 0.07	0.82 ± 0.04
NMM-HMM	0.77 ± 0.06	0.78 ± 0.08	0.69 ± 0.09
GMM-NMM-HMM	$\mathbf{0.86 \pm 0.06}$	$\mathbf{0.86 \pm 0.07}$	$\mathbf{0.87 \pm 0.04}$
VaDE [6]	0.817 (次元は 10)		

5 まとめ

本研究では, NMM-HMM の安定した学習を可能にする学習アルゴリズム GMM-NMM-HMM の提案を行なった. この手法は, 線虫の定量的な行動データ解析と全脳イメージングデータセットの組み合わせによる神経系のメカニズムの解明や, ラマンスペクトル解析による細胞種・細胞状態の推定に応用できるであろう.

参考文献

[1] Baum, L. E.; Petrie, T. The Annals of Mathematical Statistics. 1966, 37(6), pp. 1554-1563.
[2] Ciobanu, S. EPiC Series in Computing. 2021, 79, pp. 82-90.
[3] Dinh, L. et al. arXiv:1410.8516.
[4] Ghosh, A. et al. arXiv:2102.07284.
[5] Ghosh, A. et al. arXiv:2107.00730.
[6] Jiang, Z. et al. IJCAI-17. 2017, pp. 1965-1972.
[7] Kamei, K. F. et al. bioRxiv, 2023.05.09.539921.
[8] Kamimura, A; Kobayashi, T. J. Physical Review Research. 2021, 3, 033032.
[9] Kobayashi-Kirschvink, K. J. et al. Cell Systems. 2018, 7(1), pp. 104-117.
[10] Liu, D. et al. arXiv:1907.13432.
[11] Liu, D. et al. arXiv:1910.05744.
[12] Liu, M. et al. arXiv:2104.09700.