7V-07

患者への医療処置に対する強化学習適用への取り組み

上村　和貴子†　　小林　一郎†
†お茶の水女子大学

1　はじめに

集中治療室（ICU）における医療介入の管理は患者の容態を左右する最重要要素の一つである．現状，患者の容態やバイタル等の数値を元に，医師が医療介入の意思決定を行っている．本研究では，ICU における適切な医療行為を適切なタイミングで施すことを可能にする手法を検討する．具体的には，Prasad ら [1] の ICU における医療行為に対して強化学習を適用した先行研究を参考に，ガウス過程回帰を用いた，ICU データの前処理を行い，Komorowski ら [2] による ICU 患者の中でも敗血症患者に対する医療行為の課題を取り上げ実験を行い，容態改善の可能性を高める最善の方法を特定する．また，強化学習の中でも，Q-learning [3], Deep Q Network [4]，Double Deep QNetwork [5] の 3 つの手法を適用し，患者の回復に対して比較を行う．

2　強化学習を用いた医療処置

患者の容態をマルコフ決定過程としてモデル化し，強化学習を適用する．以下の様にマルコフ決定過程を定義する．

有限状態空間 $\mathcal{S}$ 時間 t における環境 (ここでは患者) の状態は s_t で表される．各時間 t での状態は，個人情報（患者の年齢，体重，性別，入院タイプ，民族性）および関連する生理学的測定値，人工呼吸器の設定，意識レベルを含む 48 次元の特徴ベクトルで表される．

行動空間 $\mathcal{A}$ 各時間 t で，エージェントは行動 (医療処置)$a_t \in \mathcal{A}$ を実行する．医療介入用の 5×5 の行動空間を設定する．静脈内（IV）液と昇圧剤（VP）の 2 つの薬剤の投与量を 0-4 の 5 段階に分け，その組み合わせで行動とする．

遷移関数 $P(s_{t+1}|s_t, a_t)$ 時間 t の状態 s_t と行動 a_t が与えられた際の次の状態への遷移確率．

報酬関数 $r(s_t, a_t) \in R$ 各時間 t での遷移後に観測された報酬．先行研究 [2] に従い，患者の全体的な健康状態の指標として，SOFA スコア（臓器不全の測定）と患者の乳酸塩レベル（敗血症患者でより高い細胞低酸素症の測定値）を用いる．SOFA スコアおよび乳酸レベルの増加にペナルティを課す．これらの指標が減少するとプラスの報酬を与える．式（1）に報酬関数を示す．また，患者の最終タイムステップで生存の場合は+15 で，それ以外の場合は-15 を与えることとする．

$$
\begin{aligned}
r(s_t, a_t) &= C_0 \mathbb{1}(s_{t+1}^{Sofa} = s_t^{Sofa} \& s_{t+1}^{Sofa} > 0) \\
&\quad + C_1(s_{t+1}^{Sofa} - s_t^{Sofa}) + C_2 tanh(s_{t+1}^{Lactate} - s_t^{Lactate}) \\
C_0 &= -0.025, C_1 = -0.125, C_2 = -0.2 \qquad (1)
\end{aligned}
$$

本研究では，強化学習の 3 つの手法，Q-learning [3], Deep Q-Network(DQN) [4]，Double Deep Q-Network(DDQN) [5] を適用し，結果を比較することで，手法ごとの医療処置に対する性能比較を行う．

3　実験

3.1　実験設定

実験データとして，医療データベースである MIMIC-III*を使用した．およそ 4 万人の患者について，個人情報，検査値，バイタルサイン，摂取/排出イベントなどの関連する生理学的パラメータを含むデータベースである．先行研究 [2] に従い，MIMIC-III より敗血症の基準を満たす患者のみを抽出し，その他の基準も準拠した． 敗血症は、SOFA スコア 2 以上で定義される臓器機能障害の証拠と感染の疑い（抗生物質の処方および微生物培養のための体液のサンプリング）の組み合わせとして定義される． 敗血症の診断は，抗生物質が先に投与された場合は 24 時間以内に微生物学的サンプルが採取されていなければならず，微生物学的サンプリングが先に行われた場合は 72 時間以内に抗生物質が投与されていなければならないという時間的基準に従った．抗生物質の投与か微生物学的サンプリングのどちらかの事象が起きた時が敗血症の発症と定義する．また，すべての患者についてベースラインの SOFA を 0 と仮定した．

MIMIC-III コホートの 80%のサンプルをモデル学習に使用し、残りの 20%をモデル検証に使用した。

A Study on Applying Reinforcement Learning to Medical Treatment on Patients
†Wakiko Kamimura(g1820107@edu.cc.ocha.ac.jp)
†Ichiro Kobayashi(koba@is.ocha.ac.jp)

*https://physionet.org/content/mimiciii/1.4/

表 1: MIMIC-III から抽出した患者の統計情報

	男女比	平均年齢	ICU 平均滞在時間 (h)	合計人数
生存者	56:44	63.4	57.6	15,583
死亡者	53:47	69.9	58.8	2,315

実際の ICU データの中身は，比較的患者に負担をかけない心拍数や呼吸数などの測定と，患者に負担がかかる動脈の pH や酸素圧等の測定のタイミングや頻度は統一されていない．これにより MDP でモデル化できないという問題が発生する．そこで，時刻が同期したデータを得るためにガウス過程回帰 [6] を前処理として行いデータの補完を行う．バイタルサインと検査結果の時間的な滑らかな相関と周期的な変動の両方をモデル化できるカーネル関数の SpectralMixtureKernel を用い，データを補完 (図 1) し，MDP でモデル化した．

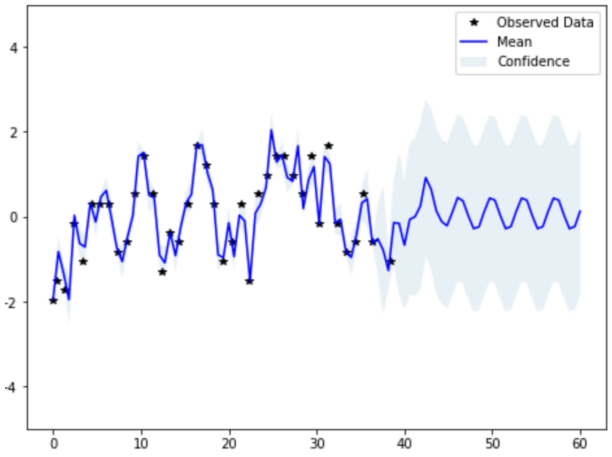

図 1: ガウス過程回帰の一例

上述した 3 つの強化学習手法を適用し、それぞれの手法による医療処置（ポリシー）の価値を検証する．実際の臨床医のポリシーによって生成された患者の軌跡と比較する OPE(off-policy evaluation) を用いて評価する．

3.2 実験結果と考察

図 2 に結果を示す．先行研究を参考に，臨床医のポリシーの「上限 95%」と今回用いた 3 つの手法のポリシーの「下限 95%」を比較することで手法の有用性を検証する．いずれの手法においても，エピソード数が 13 以内に臨床医のポリシーを超えていることが確かめられた．Q-learning と比較すると，最終的なポリシーの価値は DQN，DDQN が高いことが分かった．DQN，DDQN は最初の数エピソードで値が 0 になっているのは，試行錯誤を重ねていると見られ，報酬が得られなかったと考える．その後ある時点で急激に精度が上がったことが確認できた．また両者の手法は最終的な精度において，それほど差を示さなかった．このことに関し，DDQN は DQN の過大学習を防ぐ目的があったが，今回の実験では DQN で過大学習の問題がそれほど起きなかったと考察する．今回の実験結果においては，DQN，DDQN を用いることが良いことを示している．

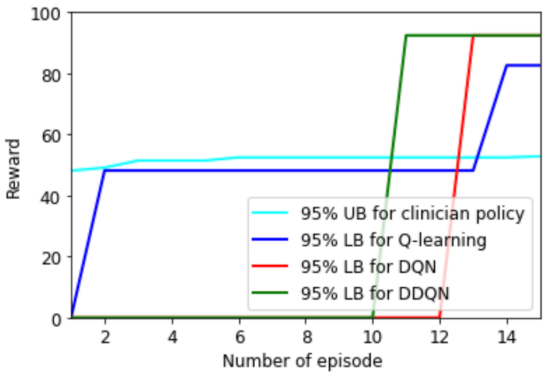

図 2: 3 つの強化学習手法による患者の回復の様子

4 まとめ

本研究では，敗血症患者に対する適切な医療介入を行うための手法の検討をした．Q-learning，DQN，DDQN の 3 つの手法で実験を行い，臨床医の医療介入との比較を行い，有用性の検証を行った．今回の実験を通じて，DQN，DDQN の手法を用いることで高い性能を出していることが確認できた．今後，さらにエピソード数を増やすことで，より一層結果の信憑性を明らかにしていくつもりである．

参考文献

[1] Niranjani Prasad, Li-Fang Cheng, Corey Chivers, Michael Draugelis, and Barbara E. Engelhardt. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *CoRR*, Vol. abs/1704.06300, , 2017.

[2] Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and AÃldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in in tensive care. *Nat Med*, Vol. 24, No. 11, pp. 1716–1720, Nov 2018.

[3] Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, Vol. 8, No. 3, pp. 279–292, May 1992.

[4] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. 2013. cite arxiv:1312.5602Comment: NIPS Deep Learning Workshop 2013.

[5] Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. 2015. cite arxiv:1509.06461Comment: AAAI 2016.

[6] Christopher Williams and Carl Rasmussen. Gaussian processes for regression. In D. Touretzky, M. C. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, Vol. 8. MIT Press, 1996.