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　(指導教員：小林 一郎)

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

医療処置に対する患者の容態の変遷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 $\mathcal{S}$　各時間 t での状態は，個人情報（患者の年齢，体重等）および関連する生理学的測定値，人工呼吸器の設定などを含む 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 $r(s_t, a_t) \in R$　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減少に正の報酬を与える．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．

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

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．

2.1　Q-learning

ある状態 s において，ある行動 a を行ったときの価値を Q 値という．状態と行動による Q 値を示された Q-Table と呼ばれる参照表を用い，この表が学習対象となる．この Q-Table に基づいて次に取るべき最適な行動を決定する．Q 値の更新を式（2）に従って行う．

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r_{t+1} + \gamma max Q(s_{t+1}, a_{t+1}) - Q(s,a)) \quad (2)$$

Q-learning は強化学習の主たる手法である．強化学習自体の有用性を示すこと，また深層強化学習との比較対象にするためにベースラインとして採用した．

2.2　DQN

Q-learning における Q-Table をディープニューラルネットワークにより近似する手法である．モデルパラメータ θ と θ^- の二つの同構造のニューラルネットワークを使用する．DQN ではニューラルネットワークの正解データの代替として以下の Target データ（3) を利用する．この Target データと実際の報酬の差である TD 誤差を最小化するようにパラメータ θ を最適化することで Q 値を更新する．

$$Target_{DQN} = r_{t+1} + \gamma max Q(s_{t+1}, argmax_a Q(s_{t+1}, a|\theta^-)|\theta^-) \quad (3)$$

Q-learning で用いる Q-Table は取りうる状態空間を全て羅列する必要がある．複雑な実世界においては状態数が多いため，テーブル関数の表現の難易度が高いという問題がある．そして DQN はその問題を改善できる．本研究においても，単純な状態空間ではないため，DQN の方が精度が高くなると考え比較対象として採用した．

2.3　DDQN

DDQN は DQN を改良したものである．学習方法は DQN と同じであり，違いは Target データの取り方にある．DDQN における Target データは式（4) で表される．DQN では，行動の選択と評価の両方に同じニューラルネットワークを利用するため，過大評価が起きることがある．DDQN では，行動価値関数に対して，価値と行動を選択するニューラルネットワークと，その行動を評価するニューラルネットワークの 2 つに役割を分ける．

$$Target_{DDQN} = r_{t+1} + \gamma max Q(s_{t+1}, argmax_a Q(s_{t+1}, a|\theta)|\theta^-) \quad (4)$$

DDQN を用いることで，過大評価の問題を改善できるため DQN より精度が高くなると考え比較対象として採用した．

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

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

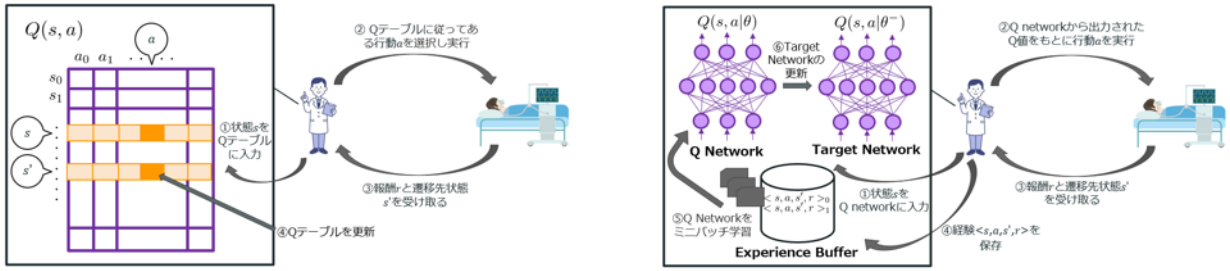

図 1: 強化学習 (左図) と深層強化学習 (右図) の概要図

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．

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，それぞれのデータ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，データを補完し，MDP でモデル化した．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

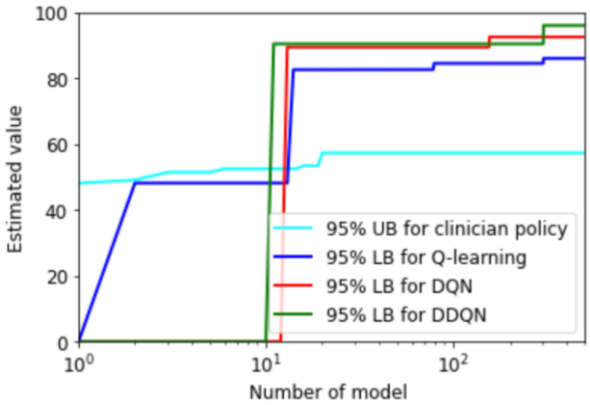

図 2: 3 つの強化学習手法による患者の回復の様子

図 2 に結果を示す．MIMIC-III で抽出した患者データのうち、80%は訓練データ，20%は検証データとして利用し，それぞれの手法で 500 モデル構築した．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，モデル数が 10 数個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．これは，3 つのポリシーは最低限の安全性を保っているということである．つまり強化学習、深層強化学習の有用性が示された．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．本研究では状態と行動空間が複雑であったため，Q-table よりもニューラルネットワークを用いる方が適しているということが示された．DQN，DDQN を比較すると，500 モデル構築後には DDQN の方が，ポリシーの価値が高いことが分かった．一方で途中経過では DQN の方が上回っている部分もあった．DDQN の提案論文 [5] における DQN と DDQN の比較では，DDQN が特に優れた性能を示すことが多いことがわかっている．また，DQN の方が優れた性能を示す場合や，ほとんど同じ性能を示す場合もあると結論付けられている．今回の実験においては特に一方が優れた性能を示したとはいえない．今回は比較のために DQN，DDQN で同様のニューラルネットワークを使った．各々最適なパラメータの探索も行うことで，精度が向上する可能性もあるため DQN と DDQN の比較についてはさらなる実験が必要である．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．

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．

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.