F-049

人間介在型強化学習における遅延のあるフィードバックに対する報酬分配
Reward Assignment for Delayed Feedback in Human-in-the-Loop Reinforcement Learning

田上 佳南太[1)]　山崎 憲一[2)]
Kanata Tanoue　Kenichi Yamazaki

1　はじめに

強化学習は，自律的な意思決定を行うための枠組みとして注目されており，ゲームやロボット制御，対話システムなど多様な応用が進んでいる．こうした応用では，学習の安全性や効率性の観点から，人間が学習過程に関与する人間介在型強化学習が注目されている．この枠組みにおいては，人間が逐次的に報酬やフィードバックを与えることで，エージェントの行動を望ましい方向へと誘導することが期待されている．

また，本研究は，推論と学習をリアルタイムで行うオンライン学習を前提とする．すなわち，エージェントは人間からのフィードバックを受け取るたびにその場でパラメータを更新し，直後の行動選択へ即時に反映する．この方式は，データ収集→オフライン学習→再デプロイを繰り返すバッチ型強化学習よりも，(i) 少量のフィードバックで漸進的に性能を向上できる，(ii) 人間側の戦略変化や環境変動に即応できる，といった利点を持つ．しかし，実環境における人間の操作には遅延が伴うことが多く，本来与えたかった行動とは異なるタイミングで報酬が記録されてしまう問題が生じる．

従来の手法では，あらかじめ固定された時間補正や，直近の数ステップに均等に報酬を割り当てるようなヒューリスティックな報酬伝搬が用いられていた．しかし，実際の人間のフィードバックには状況依存性があるため，これらの手法では，報酬を適切な行動に割り当てることが困難である．

本研究では，こうした人間介在型強化学習における遅延の問題に対処するため，遅延のあるフィードバックが「本来どの行動に対するものであったか」を補正し，その行動に報酬を割り当て直す手法を提案する．

具体的には，フィードバックが与えられた時刻と，過去の各行動の実行時刻との差（時間差 Δt）を Attention 機構の入力とし，最も関連性の高い行動を学習的に特定し，報酬を再分配する．これにより，人間の反応遅れによって発生する報酬割り当ての遅延を補正し，より正確な報酬割り当てを実現することを目指す．

2　先行研究

2.1　コンテキストバンディット学習

コンテキストバンディット（Contextual Bandit，CB）は，強化学習における決定問題の一形式であり，各試行において環境から得られる状態の代わりに，ある固定長の特徴ベクトル（コンテキスト）が観測され，その情報に基づいて最適な行動を選択することを目的とする枠組みである．CB では，各ステップで以下の手順が繰り返される．

1. エージェントは，環境からコンテキスト（特徴ベクトル）$x \in \mathcal{X}$ を観測する．
2. エージェントは，そのコンテキストに基づいてある行動 $a \in \mathcal{A}$ を選択する．
3. 環境は，その行動 a に対応する報酬 $r \in \mathbb{R}$ を返すが，他の行動に対する報酬は観測できない（バンディット設定）．

この問題における学習の目的は，期待報酬 $\mathbb{E}[r|x,a]$ を最大化するような方策 $\pi(a|x)$ を学習することである．強化学習の一般的な枠組みと異なり，CB は状態遷移や長期的な累積報酬を考慮せず，各試行が独立に扱われるため，短期的かつ即時報酬に基づく意思決定問題に適している．この性質から，CB はインタラクティブなシステムにおけるユーザからのフィードバックを用いたオンライン学習や，推薦システム，広告最適化などの応用で広く用いられている．

上記の CB の枠組みは，人間からのインタラクティブなフィードバックを活用した学習においても有効であり，Suhr ら [1] はこれを自然言語指示に従うエージェントに適用した．彼らの手法では，各ステップを以下の三要素で構成する．

- コンテキスト x：ユーザからの自然言語による指示と，エージェントの観測情報（視覚・位置・履歴など）からなるベクトル
- 行動 a：エージェントが実行する離散的な原始行動（例：前進，回転，カードの選択など）
- 報酬 r：行動実行後 0.2 秒以内に受け取った人間の正または負のフィードバックがあった場合に，+1 または -1 として割り当てる．もし該当ステップにフィードバックが無い場合は，直後から最大 8 ステップの範囲で最初に観測されたフィードバックを遡って割り当てるヒューリスティックな遅延補正を行う

学習には割引率を 0 とした方策勾配法を用い，過去の行動選択確率を考慮した逆傾向重み（Inverse Propensity Score，IPS）を導入している．これは，過去の行動選択方策と現在の方策の差異に起因するバイアスを補正するためである．この手法の特徴は，報酬関数を明示的に設計することなく，人間のフィードバックによって指示追従行動を学習できる点にある．また，フィードバックはオンラインで逐次受け取られ，方策は継続的に更新されるため，少ない試行回数でも性能の改善が可能である．

1）芝浦工業大学大学院理工学研究科電気電子情報工学専攻 Graduate School of Engineering and Science, Electrical Engineering and Computer Science, Shibaura Institute of Technology
2）芝浦工業大学大学院理工学研究科電気電子情報工学専攻 Graduate School of Engineering and Science, Electrical Engineering and Computer Science, Shibaura Institute of Technology

2.2 技術課題

この手法では報酬の割り当てに関していくつかの課題が残されている．最大の問題は，ユーザがフィードバックを送るタイミングに遅延が存在することである．[1] では，前節で述べた「0．2 秒以内のフィードバック検出＋最大 8 ステップ遡及」というルールベースで報酬を割り当てている．

このような処理は，すべての状況において適切とは限らず，フィードバックが本来意図していた行動と異なる行動に報酬が割り当てられる可能性がある．さらに，報酬伝搬の設計が「0．2 秒」「8 ステップ」という固定閾値を予め定義するだけのルールベース手法であるため，学習の初期段階やノイズの多い環境では報酬割り当ての精度が不十分となる．このことが，学習の不安定化や過学習につながる恐れがある．

したがって，ユーザのフィードバックが「本来どの行動に向けられていたのか」を，タイミングと行動内容の情報に基づいて学習的に推定・補正する仕組みが求められる．

2.3 関連研究との比較

報酬の遅延やスパース性に対処する代表的手法としては RUDDER[2] が挙げられる．RUDDER はエピソード全体（または固定長チャンク）を入力とし，LSTM で総報酬を予測したうえで，その予測値の差分から各タイムステップの貢献度を推定し，報酬を再配分することで Q 値推定を単純化する．

しかし RUDDER は

- エピソード終了後にまとめて学習を行うバッチ型である
- フィードバックを一時的に蓄積し，後処理で再配分するためリアルタイムの方策更新ができない

という設計上，逐次的にフィードバックを受け取り即時に学習を反映させる本研究のオンライン学習設定とは相性が悪い．

3 提案手法

本研究では，人間介在型強化学習において，ユーザが遅延を伴って送信するフィードバックを，本来意図していた行動に適切に割り当て直す仕組みの実現を目的とする．従来手法のような固定遅延補正やヒューリスティックな報酬伝搬に頼らず，行動と報酬の関係を学習的に推定することで，フィードバックの遅延に対応できる報酬分配を目指す．

提案手法では，報酬が与えられた時刻と，過去の各行動の実行時刻との差（時間差 Δt）を計算し，その情報と行動の内容，さらに行動時点での環境状態を入力として自己注意機構（Attention）に与える．Attention は，報酬が「どの行動に向けられたものか」を学習的に判断し，得られたスコアをもとに報酬を過去の行動に重み付きで分配する．図 1 は，ユーザのフィードバックが遅れて到着する場合に，過去の行動列に対してどのように時間差と Attention スコアを用いて報酬が分配されるかを示した例である．

具体的には，報酬イベントを表す固定の Query ベクトル q_0(学習可能) と，各行動に対応する Key ベクトルを用いて Attention スコアを計算する．Key は，行動の埋め込み e_a，時間差 Δt，および環境状態の埋め込み e_s を結合する．Value には e_a を用いる．このスコア列と人間のバイナリフィードバックの積を報酬として，各行動に分配する．

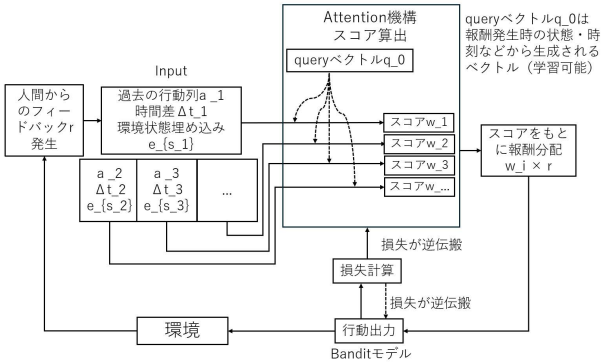

図 1 提案手法の流れ

分配された報酬は，方策やバンディットモデルの学習に利用される．このとき，分配結果に基づく報酬の増減が Attention スコアの生成元にも逆伝播され，Attention 機構自体も報酬最大化のために更新される．これにより，報酬の遅延が一律でない状況や環境文脈に応じて，状況適応的に報酬の割り当て先を補正できることが期待される．

評価実験では，Suhr らが構築した逐次的指示追従環境を用い，提案手法が従来の報酬伝搬手法と比較して，より高い報酬割り当て精度や学習効率を実現できるかを検証する．

4 おわりに

本研究では，人間介在型強化学習におけるフィードバックの遅延問題に対処するため，遅れて到着した報酬を本来意図された行動に割り当て直す仕組みとして，Attention 機構に基づく報酬再分配手法を提案した．従来のルールベースによる遅延補正では困難であった状況依存性や非一様な遅延に対して，本手法は柔軟に対応可能であり，より適切な報酬割り当てを実現できることが期待される．今後は，逐次的な指示追従タスクを対象とした実験を通じて，報酬割り当ての精度や学習効率への効果を定量的に検証し，提案手法の有効性を明らかにする予定である．

参考文献

[1]Alane Suhr, Yoav Artzi: “Continual Learning for Instruction Following from Realtime Feedback”, Advances in Neural Information Processing Systems, Vol. 36, pp. 32340–32359(2023).

[2]Jose A. Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, Sepp Hochreiter: “RUDDER:Return Decomposition for Delayed Rewards”, Advances in Neural Information Processing Systems, Vol. 32, (2019).