F-039

シナリオごとの Q 値分解に基づく強化学習の解釈性向上に関する一検討
A Study on Interpretation of Reinforcement Learning Based on Q-Value Decomposition by Expected Scenarios

土屋 祐太† 森 靖英† 恵木 正史†
Yuta Tsuchiya Yasuhide Mori Masashi Egi

1. はじめに

強化学習は、行動により報酬が与えられる環境において、報酬を最大化する行動が出力されるように、機械学習モデル (AI) のパラメータを学習していく仕組みである。その高い予測性能から、社会インフラや医療現場などの業務にまで適用範囲を広げることが期待されている。例えば、予想される自然災害による被害を最小限に抑えるために、あらかじめ人員などのリソースを適切に配置する事前対策計画を立案することが挙げられる[1]。

しかし、このようなミッションクリティカルな業務に機械学習システムを活用するためには、高い有用性に加え、透明性、公平性、解釈性など様々な性質についての要件を満たすことが求められる。そこで、機械学習システムの判断根拠を説明する技術である XAI (eXplainable AI) の研究が急速に進展している。例えば、AI に入力した画像特徴量において、AI が重要視した部分をヒートマップで可視化する手法が挙げられる[2]。教師あり学習の枠組みでは、先のような特徴量に対する説明技術の開発が盛んに行われている。一方で、強化学習 AI の行動は、将来得られる報酬や事象を見越して学習されるものであるため、特徴量に対する「過去向きの説明」ではなく、AI が意図した将来の事象に対する「未来向きの説明」が求められる。

そこで、AI が行動を選択するために想定した将来の状態(シナリオ)を説明する手法として、最も起こる確率の高いシナリオ、あるいは起こりうる全ての状態と行動に対して期待される報酬のテーブルを、説明のシナリオとして用いる方法が提案されている[3][4]。しかし、前者は AI の意図を説明するために十分な情報を得ることができず、利用者の様々な関心に対応した説明をすることは困難である。後者は網羅的に記述できる反面、状態の組み合わせが多い複雑な問題への適用が難しいという課題があった。

そこで本稿では、Actor-Critic 型の強化学習 AI を想定し、利用者が注目する状態ごとに報酬の期待値 (行動価値) を推定することで、複雑な問題に対してもシナリオを説明可能にする手法を提案する。従来の強化学習 AI が有する、全ての状態に対しての行動価値を評価する Critic (以下、基準 Critic) に加えて、利用者が指定した状態ごとに行動価値を評価する Critic (以下、説明 Critic) を導入し、基準 Critic と説明 Critic がめざす真の行動価値を揃えて説明 Critic を学習させる。それによって、利用者が注目したい任意の状態に対してのみ行動価値を評価して説明することができ、状態の組み合わせが多い複雑な問題に対しても想定シナリオを説明可能となる。さらに、提案手法の検証のため、複数の状態の組み合わせを有する計画問題として、簡易的な電力系統の災害リスク対策のためのリソース配置最適化問題による実験を行った。その結果、本手法によって妥当な説明を提示できることを確認した。

†(株)日立製作所
〒185-8601 東京都国分寺市東恋ヶ窪 1-280
Hitachi, Ltd.
1-280, Higashi-koigakubo, Kokubunzi, Tokyo 185-8601, Japan

2. 強化学習 AI の概要と解釈性向上手法

2.1 強化学習 AI

本論文は、強化学習 AI として Actor-Critic 型のモデルを想定している[5]。これは、Q 値を正確に推定するように学習する Critic と、行動選択の確率分布や行動自体を出力する Actor の 2 つのニューラルネットワークを用いた手法であり、連続値制御への適用も含め幅広く応用されている。Q 値とは、ある状態における行動の期待報酬 (行動価値) を表すスカラー値であり、以下の式で表される。

$$Q(s_t, a_t) = \mathbb{E}_{s_{t+1}\sim P(s_{t+1}|s_t,a_t)}\left[R_t + \gamma \max_{a_{t+1}} Q_{EX}(s_{t+1}, a_{t+1})\right] \quad (1)$$

ここで、s_t, a_t, R_tはそれぞれ時刻tにおける状態、行動、報酬を表し、Q_{EX}は Critic が出力した次の時刻の Q 値、γは学習率、$P(s_{t+1}|s_t, a_t)$は状態遷移関数を表す。

Critic から出力される Q 値は、将来の期待報酬である行動価値を表すことから、AI が想定したシナリオを把握するための情報を持つと期待できる。しかし、式(1)の Q 値は複数の状態遷移の期待値であるため、このままでは想定された具体的な将来のシナリオを解釈することができない。そこで本稿では、利用者が注目する状態を指定し、その状態ごとに行動価値を評価する Critic (説明 Critic) を追加することで、想定されたシナリオを説明する手法を提案する。

2.2 シナリオごとの Q 値分解に基づく説明手法

提案する説明手法の構成概要を図 1 に示す。説明 Critic は、Q 値を評価する元の Critic (基準 Critic) と同様の入出力を持つニューラルネットワークであり、利用者の注目するシナリオの数だけ予め生成される。説明 Critic が推定する行動価値は、式(1)の状態遷移関数Pにおいて各説明 Critic に対応するシナリオを固定した場合の、基準 Critic の Q 値となる。それによって利用者は、AI の行動が具体的にどのようなシナリオにおいて有効か、もしくはリスクがあるかを解釈することができる。さらに、AI の行動と利用者が指定する行動を、シナリオとともに比較することもできる。

説明 Critic の学習は、その説明対象となる強化学習 AI の学習と同時もしくは完了した後に行われ、それぞれ対応する状態のデータのみを学習する。そのため、エピソードを繰り返して学習データを蓄積する際に、各時間ステップで発生した状態遷移と対応する説明 Critic ごとに、別々のラベルが学習データへ付与される。

ここで、基準 Critic は、以下の Temporal Difference (TD) 誤差の最小化を目的として学習を行っている。

$$E(s_t, a_t) = \left(R_t + \gamma \max_{a_{t+1}} Q_{EX\_tar}(s_{t+1}, a_{t+1}) - Q_{EX}(s_t, a_t) \right)^2 \quad (2)$$

Q_{EX_tar}は、パラメータ固定されたターゲットネットワークによる Q 値を表す。

従来、複数の Critic を用いた強化学習手法は、複数タスクへの適用などの目的で提案されている[6]。しかしこれらは、Critic ごとに別々のターゲットネットワークを用いて学習しているため、本手法にそのまま適用すると、説明 Critic の Q 値が基準 Critic の Q 値から遠ざかってしまう。そこで、説明 Critic における式(2)の TD 誤差計算に用いるターゲットネットワークを、基準 Critic と同様にする。それによって、基準 Critic と説明 Critic がめざす真の行動価値を揃えて説明 Critic を学習することができるようになり、基準 Critic と一貫性を持った Q 値が推定可能となる。

3. 数値実験

複数の状態の組み合わせを持つ問題において、提案手法による説明の妥当性を検証するため、電力系統の災害リスク対策を目的としたリソース配置問題による実験を行った。Actor は状態として 3 つのエリアの「停電量・リソース配置・時刻」を受け取り、行動として「集中・分散」のリソース配置を指定できる。各エリアの配置が終了した段階で、表 1 に示す 4 種類の状態遷移のいずれかが発生する。遷移先で集中型の事故が発生し、行動として集中配置を選んでいた場合、AI に報酬が与えられる。分散配置についても同様である。基準 Critic は全ての状態に対する Q 値、説明 Critic は表 1 の状態それぞれに対する Q 値を出力する。ここで、説明の妥当性を①基準 Critic の Q 値が説明 Critic の最大・最小 Q 値の間に存在、②行動に対して報酬の高いシナリオの Q 値が高い、の 2 つの観点で検証した。

1000 エピソードで学習を行い、Actor が出力した行動に対する、基準 Critic と説明 Critic から得られた各状態に対する Q 値を図 2 に示す。ここで、Actor は事故②に対して報酬が高い分散配置を行った。図 2 を確認すると、事故②に対する Q 値が最も高くなっていることから、説明の妥当性②の観点を満たしている。さらに、基準 Critic の Q 値は、説明 Critic による Q 値ベクトルの最大・最小の間に位置しており、説明の妥当性①を満たすため、説明 Critic が提示する Q 値とシナリオの妥当性が確かめられた。

さらに、なぜ集中配置は適切でないのかを説明するため、利用者が行動として集中配置を指定した場合の Q 値を図 3 に示す。図 2 と図 3 より、集中配置は集中型の事故①と③に対する Q 値が高い一方で、分散配置よりも事故②と事故なしに対する Q 値が大きく下がっているため、基準 Critic の値は分散配置の方が大きくなっている。以上より、行動の違いを具体的なシナリオとともに比較することができた。

4. おわりに

本稿では、強化学習 AI が出力した行動に対して、AI が想定した具体的なシナリオからその行動意図を解釈することを目的として、シナリオごとに Q 値を分解して評価する手法を提案した。利用者の注目に基づいて分解することから、複雑な問題に対しても説明可能となり、さらにめざす真の行動価値を基準 Critic と揃えることで、一貫性を保った説明を生成することができた。

今後は、より複雑な問題における検証や、学習に必要な計算量を削減するための方法について検討していく。

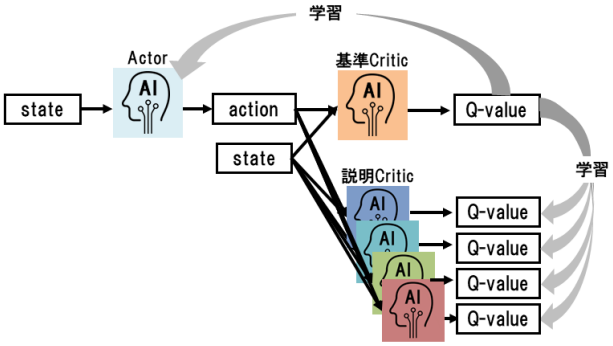

図 2　提案手法の概要

表 1　状態遷移の組み合わせ

	エリア 1	エリア 2	エリア 3	遷移確率
事故①	集中	集中	なし	10 %
事故②	分散	分散	分散	40 %
事故③	なし	なし	集中	10 %
事故なし	なし	なし	なし	40 %

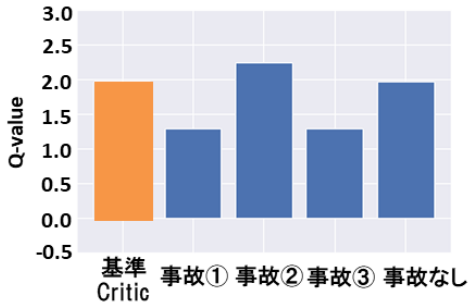

図 2　「分散」配置に対する Q 値

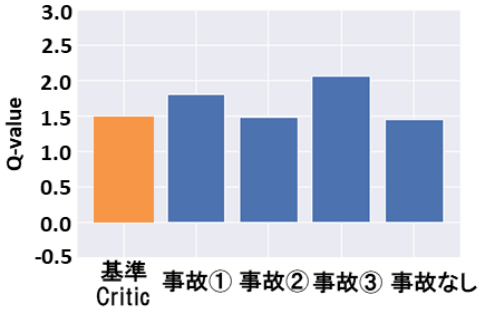

図 3　「集中」配置に対する Q 値

参考文献

[1] M. M. Hosseini and M. Parvania, "Resilient Operation of Distribution Grids Using Deep Reinforcement Learning", IEEE Transactions on Industrial Informatics, Vol. 18, No. 3, pp. 2100-2109 (2022).
[2] S. Greydanus, A. Koul, J. Dodge, and Alan Fern, “Visualizing and Understanding Atari Agents,” Proceedings of the 35th International Conference on Machine Learning, PMLR 80, (2018).
[3] J. V. D. Waa, J.V. Diggelen, K.V.D. Bosch, and M. Neerincx, “Contrastive Explanations for Reinforcement Learning in terms of Expected Consequences,” arXiv, 1807.08706, (2018).
[4] H. Yau, C. Russell, and S. Hadfield, “What Did You Think Would Happen? Explaining Agent Behaviour through Intended Outcomes,” Workshop on Extending Explainable AI Beyond Deep Models and Classifiers, PMLR 119, (2020).
[5] V. R. Konda and J. N. Tsitsiklis: “Actor-Critic Algorithms”, Society for Industrial and Applied Mathematics, vol. 42, (2001).
[6] S. Mysore, G. Cheng, Y. Zhao, K. Saenko, and M. Wu, “Multi-Critic Actor Learning: Teaching RL Policies to Act with Style”, International Conference on Learning Representations, ICLR (2022).