Offline Reinforcement Learning from Imperfect Human Guidance(Abstract_要旨)

AUTHOR(S):

Zhang, Guoxi

CITATION:

Zhang, Guoxi. Offline Reinforcement Learning from Imperfect Human Guidance. 京都大学, 2023, 博士(情報学)

ISSUE DATE:

2023-07-24

URL:

https://doi.org/10.14989/doctor.k24856

（続紙 1）

京都大学	博士（情報学）	氏名	Guoxi Zhang
論文題目	Offline Reinforcement Learning from Imperfect Human Guidance (不完全な人間の誘導からのオフライン強化学習)		

（論文内容の要旨）

本論文は、逐次的な意思決定問題である強化学習において、人間によって与えられた質にばらつきのある教示から学習する際に想定される問題を解決するためのアプローチについての研究結果をまとめたものであり、全６章から構成されている。

第１章は序論であり、本論文の目的とその内容を概観している。強化学習は一連の意思決定の結果に対する報酬関数を通じて、意志決定方策を学習する一般的な枠組みであるが、適切な報酬関数の設計がボトルネックであった。この問題に対して、人間による例示や事例の評価データを事前に収集し、これらを利用するアプローチが注目を浴びている。しかしながら、データを与える人間によって質や傾向にバラつきが生じることや、学習された報酬関数には解釈性が乏しいこと、また、人間による例示には異なる意図に基づくものが混在していることなどの問題があった。本章ではこれらの問題に対する種々のアプローチを検討するという展望について述べている。

第２章では、準備として、本論文で共通して用いられる理論的枠組みとなる、強化学習の問題設定と、本論文で提案する手法の基礎となる、方策学習手法である Actor-Critic 法に基づく種々の深層強化学習アルゴリズムおよび、人間によって与えられた選好にもとづく強化学習を行う際の基本的なモデルとなる Bradley-Terry モデルなどについて紹介している。

第３章では、多数の人間が行動履歴の好ましさを相対評価したデータから、強化学習の報酬関数を学習する問題を扱っている。多数の人間が関わることで、大量のデータを評価できる一方、人によって評価の質や傾向のバラつきが問題となるが、本章ではこの問題を、それぞれの人間の信頼度として明示的に報酬関数と評価モデルの中に取り込んだ深層強化学習手法を提案することで解決している。複数のシミュレーション課題を用いた実験では、提案手法が一貫して従来手法よりも高い性能を示すことを確認している。

第４章では、強化学習の説明可能性について論じている。近年の、AI の信頼性への関心の高まりによって、説明可能 AI(XAI)が議論されているが、強化学習における XAI 研究は殆ど手付かずである。本章では、報酬関数の中に、強化学習エージェントが行動履歴のどの箇所に着目しているかを示す注意機構を組み込むことで、根拠となる箇所を特定し視覚化する方法を提案している。複数のシミュレーション課題を用いた実験では、提案手法によって、性能を殆ど犠牲にすることなく、重要な場面を特定できることを確認している。

第５章では、一連の意思決定が人間による例示として与えられる状況での強化学習問題を扱っている。異なる人間によって、あるいは異なる意図をもって与えられた例示が混在している状況では、これらを区別せずに扱うと、学習に失敗することがある。本章では、この状況を複数の方策の混合分布を用いてモデル化することで解決を図る深層強化学習法を提案している。複数のシミュレーション課題を用いた実験では、複数の方策が混在していると考えられる状況において、提案手法が従来手法よりも高い性能を示

すことを確認している。

第６章は結論であり、本論文で得られた成果を要約している。即ち本論文は、逐次的な意思決定問題である強化学習において、多数の人間によって与えられた、質が均一でない教示をもとに、高い性能を実現し、ときには解釈性を兼ねそろえた報酬関数や方策を獲得する方法を提案したものであり、その有効性を実験的に示したものである。本章では最後に、将来の課題・展望として、できるだけ低コストで有用なデータを獲得するための能動学習法との融合や、複数の価値観が混在する複雑な効用関数のデザイン、マルチモーダルデータの扱いなどを挙げ、本論文を結んでいる。

（続紙 ２ ）

（論文審査の結果の要旨）

本論文は、強化学習において、人間によって与えられた質にばらつきのある教示から学習をおこなう際に想定される問題の解決に取り組んだものであり、得られた主な成果は次の通りである。

１．多数の人間による相対評価が与えられた行動履歴データから、人による評価の質や傾向のバラつきを明示的に報酬関数と評価モデルの中に取り込むことで、報酬関数を高い精度で学習する深層強化学習手法を提案するとともに、その性能を実験的に確認した。

２．報酬関数の中に、強化学習エージェントが行動履歴のどの箇所に着目しているかを示す注意機構を組み込むことで、根拠となる箇所を特定し視覚化できる、説明可能性を持った強化学習法を提案し、その性能を実験的に確認した。

３．異なる人間によって、あるいは異なる意図をもって与えられた例示が混在している状況を、複数の方策の混合分布を用いてモデル化することによって、一連の意思決定が人間による例示として与えられる状況に適した深層強化学習法を提案し、その性能を実験的に確認した。

以上、本論文は、逐次的な意思決定問題である強化学習において、多数の人間によって与えられた、質が均一でない教示をもとに、高い性能を実現し、ときには解釈性を兼ねそろえた報酬関数や方策を獲得する方法を提案し、その有効性を実験的に示したものであり、学術上・実応用上寄与するところが少なくない。よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、令和５年６月２３日に実施した論文内容とそれに関連した口頭試問の結果、合格と認めた。なお、インターネットでの全文公表を行うことについて支障がないことを確認した。