論文審査要旨

論文審査委員会

主査　大和　淳司　　副査　田中　久弥

副査　真鍋　義文　　副査　竹川　高志

副査　鮫島　和行（玉川大学）

論文題目：　冗長な観測に対する状態削減問題と目的指向な強化学習環境の推定

学位申請者：　高橋 春輝（工学研究科情報学専攻博士後期課程　ED21501）

本論文は、人工知能分野における強化学習エージェントが複雑で膨大な情報量を持つ観測を前提とした環境で直面する問題に対する解決手法を提案し、評価した成果をまとめたものである。

強化学習は、観測・報酬を提供する環境内で長期報酬を最大化するように学習するエージェントについて研究する分野である。強化学習エージェントの性能は深層学習の活用などにより急速に向上しており、対戦ゲームやビデオゲームにおいて人間を凌駕する性能を発揮するようになっている。しかし、これらの性能を得るために、既存の大規模データやオフラインでのシミュレーションを前提としており、実課題に対応するためにオンラインでの学習効率が求められている。また、学習結果の解釈が困難で実社会での利用についてはリスクが生じるという問題点が残されている。一方、計算論的神経科学分野を中心に、行動による環境の推移などを陽にモデル化し推論するモデルベース強化学習の取り組みも活発に行われており、これらの研究は人間の持つ柔軟な環境理解と意思決定能力の理解とも関連が深い。これらの研究を踏まえ、強化学習を新たな枠組みで統合した形で整理することがさらなる研究の発展に有用である。

モデルベース強化学習では、部分観測マルコフ決定過程（Partialy Observed Markov Decision Process; POMDP）と呼ばれる枠組みで議論されることが多い。申請者は、POMDPは複雑な環境での強化学習に対しては部分的な問題しか扱えていないという課題を発見し、本論文においてその課題を解消をする枠組みを状態削減問題として再定義している。状態削減問題は、観測・行動・報酬に関する過去の履歴を元に、現在の状況（状態）を報酬予測に必要なシンプルな形で推定する。さらに、状態削減問題は POMDP の扱う部分観測問題に加えて、画像からの物体認識などの情報の構造に依存する冗長性と報酬に関連しない情報を含む冗長性に分解して考えることができることを示している。

申請者はこれらの考察を元に、情報の構造に依存する冗長性を削減する手法の一つとして判別的ラプラシアン固有マップ法（Laplacian Eigenmap）を、報酬に関連しない情報を削減する手法として目的指向環境推定 （Goal Oriented Environment Inference; GOEI）を提案し、本論文において提案手法を評価した成果をまとめている。

本論文は、６章で構成されている。１章では、序論として、計算論的神経科学や人工知能分野におけるエージェントに関する研究を概観することで状態削減問題を導出するにあたった経緯を説明している。特に、実世界でのエージェントの実現において課題となる膨大な情報から有益な情報を抽出する情報処理と目的を達成するための行動制御に焦点を当てて整理しており、エージェントの課題である説明可能性が状態削減問題を解くことで向上する点について整理している。

第２章では、未知の環境ダイナミクスとエージェントの相互作用による強化学習環境において、既存研究のエージェントの取り組みを整理し、状態削減問題を定義している。状態削減問題において、報酬を予測するために必要な最小の状態集合をコアと定義し、コアに基づく環境ダイナミクスを推定できれば最適戦略の学習にかかる試行錯誤を最も減らすことができることを示している。また、状態に対する観測の冗長性を分類し、既存の取り組みとの対応関係を整理している。

第３章では、観測の構造を利用して状態削減問題にアプローチする次元削減による手法を提案し、評価している。具体的には、真のコアが既知であると解釈できる線形判別分析を、コアの予測値である観測値間の類似度で近似した手法が、類似度に付加されたノイズに対して頑健にコアを推定できることを示している。

第４章では、モデルベース強化学習でよく用いられる POMDP に対して、第２章の考察を元に、報酬に依存しない情報が観測に含まれる冗長的観測 MDP（Redundantly Observable MDP; ROMDP）を定義し、POMDP および ROMDP と状態削減問題との関係を整理した上で、POMDP に対するエージェントモデルである完全環境推定（Complete Environment Inference; CEI）と目的指向環境推定（GOEI）の導出と実装を行っている。

第５章では、４章の実装に基づき、シミュレーション環境を用いて、GOEI により CEI では不可能な報酬に無関係な情報を削減することによる状態削減が実現しており、さらに状態削減により学習に必要なサンプルサイズや非定常性に対する応答性の改善が実現できていることを確認している。

最後に、第６章において全体のまとめを行い、今後の展望として構造依存の情報の削減とGOEIをどのように組み合わせるか、ROMDP と POMDP を組み合わせたより汎用的なエージェントへの展開について考察を行っている。

本研究は、現在主流となっている強化学習のアルゴリズムに対して、新たな視点での問題提起とその解決方法を示しているという点で画期的な研究である。単なる既存手法の部分改良ではなく本質的な改善をもたらしていると言える。新たな視点からの検証に特化しているという点で具体的な応用という観点からは限定的な結果だが、他の手法との柔軟に組み合わせることができることから発展性が高く、今後の多くの研究の基盤となる。

以上より、本論文は、博士（情報学）の学位論文に値するものと認める。

2024年7月11日

論文所見

玉川大学　鮫島 和行

論文題目：　冗長な観測に対する状態削減問題と目的指向な強化学習環境の推定
学位申請者：　高橋　春輝 氏（工学研究科情報学専攻博士後期課程　ED21501）

本論文は、人工知能の根幹問題である環境状態推定問題について、一定の仮定のもとの最適行動を学習する問題（強化学習）の枠組みで定式化し、従来提案されてきた解決方法とこの論文で提案する手法の対比を、理論と数値実験の双方で検証した研究である。古典的にはフレーム問題とよばれるこの問題は、現代的な機械学習による人工知能研究では部分マルコフ決定過程(Partially Observable Markov Decision Process; POMDP)として定式化され、環境内の情報からいかにして最適な行動に必要な観測から状態への写像を得るのかについての問題として扱われてきた。

従来の手法では、隠れた状態は現在の観測だけでは特定できず過去の行動・報酬・観測から推定する問題となり非常に困難な問題となる。一方高橋氏は、現在の状態は現在の観測からえられる情報から推定できるが、冗長な情報、特に報酬を最大化するのに関連のない情報を含むものから、必要とされるコア状態を推定する問題として定式化されるRedundantly Observable Markov Decision Process; ROMDPであるとして、報酬に関連するコアな状態を推定する問題を扱っていることを明示的に述べている。ROMDPではPOMDPで想定される推移則とは異なる生成モデルを設定することにより、従来とは異なる視点で定式化した環境状態推定問題であり、そのアイデアは独創的である。本来実環境で動くロボットなどでは観測が制限されていることを仮定しているが、観測が十分にされているにも関わらず、その中から重要な観測を選べない問題に置き換えて考えることもできるという逆転の発想がある、この問題提起は重要であり学術的価値が高い。

POMDPとROMDPの違いを理論的に示した上で、それぞれの観測から状態を推定する手法CEIとGOEIという変分ベイズ推論とノンパラメットリック法を用いた手法による推定アルゴリズムを提案し、その実装方法と特徴を考察している。CEIは報酬の推定だけでなく状態予測に基づく方法で報酬に関連のないものまで推定しようとするのに対して、GOEIでは報酬に関連する状態のみを推定しようとする原理を説明した。そのため、より小さな状態推定を得やすく、高速に最適行動を学習できる可能性について述べている。また、GOEIから探索と搾取のバランスを自然にとることができるATS手法も提案している。

提案手法の検証として、コアな状態遷移に直交した行動に依存した状態遷移を行うノイズ状態と報酬に依存して遷移するノイズ状態の直積空間内の点として観測値を与えた場合についての数値実験を行い、提案手法であるGOEIとCEIの違いについて結果を

示しており、従来の状態から観測が生成されるというモデルを用いて推定を行うCEIよりも、より小さな状態集合の遷移として推定可能で、学習もCEIより高速に行うことができることを実証している。これらの結果は、このような小さな問題を超えて、実世界の強化学習エージェントに応用可能な実用的アルゴリズムにつながることが期待される。

博士論文公開発表会において、高橋氏は上記のROMDPの問題からCEIおよびGOEIのアルゴリズムの基礎となる数理について、数式を使わずに、本質的な概念とベイジアンネットワーク図による説明を用いて明快に説明し、アイデアの根幹である、観測が隠れた状態から生成される確率モデルではなく、現在の観測の中から状態が生成されるように視点を変えることで問題を解決する手法の本質として明快に説明しており、研究者としてのプレゼンテーションの能力も高いことを示した。

以上のことから、博士学位論文及び発表会の内容は博士(情報学)の学位に相応しいと判断できる。