博士論文要旨

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

工学院大学大学院工学研究科

情報学専攻　博士後期課程　高橋春輝

本論文は，実世界で意思決定を行うエージェントが直面する膨大な情報から有益な情報を抽出する状態削減問題に関する論文である．この問題を異なるアプローチで解決する 2 つの手法を提案し [1, 2]，これらの提案手法の性能評価を行なった成果をまとめたものである．

人間や動物は，感覚情報を通じて世界を認識し，試行錯誤を行いながら学習を行い，段階的に適切な行動を行うことができるようになる．特に人間は視覚情報を効果的に処理し、目的を達成するのに有益な情報を抽出しながら，長期的な手順の計画を必要とするタスクを解決することができる．このような高度な知的能力を持つエージェントが実現できれば、危険の伴うタスクや高精度な作業を自動化することが可能となる．

実世界のタスクに応用できるエージェントを実現するには，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の 2 つの技術的課題がある．そして，このようなエージェントは，なるべく少ない試行錯誤でタスクに適応できるようなサンプル効率の条件を満たすことが望ましい．また，エージェントの利用者が安心して使用するためには，エージェントの行動原理に対する説明可能性が応用時に課題となる．

計算論的神経科学や人工知能の分野では，これらの技術的課題を解決するのに有用な手法が研究されてきた．Deep learing による情報処理と強化学習による行動制御を組み合わせた深層強化学習によるアプローチは，有益な情報に対する因果構造をブラックボックス化することで，幅広いタスクに応用することができる．しかし，因果構造をブラックボックス化したことで，学習に多くの試行錯誤を必要とするだけでなく，行動原理が不透明となるため，実世界でのエージェントの実現が困難である．

これに対し，全ての情報の因果構造を直接モデル化するベイズ推論を情報処理に代用することで，学習に必要な試行錯誤を軽減することができる．また，因果構造をモデル化することで，エージェントの行動が将来に与える影響を予測できようになるため，説明可能性も高い．しかし，このアプローチでは，全ての情報を考慮するため，膨大な情報を処理するのに多くのメモリを要求する実装上の課題が残っている．

実世界における情報には，エージェントが目的を達成するのに有益な情報だけでなく，それとは無関係な情報も多く含まれる．そのため，膨大なから有益な情報のみを抽出し，それに対する因果構造をモデル化することができれば，メモリ要求の課題は軽減する．加えて，有益な情報に限定した簡潔な因果構造は，人にとって理解がしやすくなる利点もある．

そこで，本論文では，膨大な情報から有益な情報に関する因果構造を推定する問題を状態削減問題として定義する．そして，この問題を解く手法として，情報の構造を利用した次元削減によるアプローチと情報の生成過程を利用した環境推定によるアプローチを提案する．次元削減によるアプローチは膨大な情報を持つ観測値を低次元で表現する状態削減を可能とし，環境推定によるアプローチは時系列の因果構造を考慮してさらに有益な情報を抽出する状態削減を可能にする．実験では，これらのアプローチが有益な情報を損失せずに状態を削減できることと，その状態削減が行動制御に与える影響について検証した．

以下に，各章の概要を示した．

1 章　はじめに

1 章では，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た．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た．また，エージェントの利用時における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も整理した．

2 章　強化学習環境に対する状態削減問題としての定式化

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，膨大な情報から有益な情報を抽出する過程を強化学習環境での学習に組み込んだ状態削減問題を定義した．強化学習環境において，エージェントの目的は将来に獲得する総報酬を最大化するように，情報となる観測値から行動を制御することである．将来の総報酬を最大化するような環境に適応できる最適な戦略を学習するには，環境を探索することで様々な状況を経験する必要がある．しかし，環境を探索しすぎれば報酬獲得の機会を逃すことになるため，少ない探索回数で最適な戦略を学習できるような行動制御が望ましい．このような行動制御を行うために，観測値や報酬そしてエージェントの行動の履歴を手がかりにして，未来の報酬を予測できる環境ダイナミクスを推定する問題を状態削減問題として定義した．この問題の解となる，報酬予測に寄与する情報を持った状態集合は，一意に定まらないことを確認し，そのような解の中で最小の状態集合をコアと定義した．環境ダイナミクスを構成する状態数が多くなるほど，最適戦略を学習するのに必要な試行錯誤が増えるため，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．

実世界でのエージェントの実現を想定して，1) 遷移則や報酬則が変化する非定常な環境での戦略学習の性能，2) 目的に基づいた状態削減，3) 環境ダイナミクスの推定に基づいた説明可能性の 3 つの基準から関連手法を整理した．それぞれの関連手法によるエージェントはいずれかの基準に特化しているため，研究目的としてこれら 3 つの基準を満たすエージェント開発の必要性を示した．

3 章　次元削減によるアプローチ

3 章は [1] の内容であり，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する．次元削減は，高次元データから低次元データへの線形変換として一般に定義される．このとき，高次元データを観測値集合とし，低次元データをコア集合とみなすことで，観測値が持つ情報の構造を低次元で表現できるという制約下での状態削減問題を解くことができる．

次元削減によるアプローチとして，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判別的ラプラシアン固有マップ法を提案した．結果として，この手法は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が示された．

4 章　環境推論によるアプローチ

4 章は [2] の理論部分の内容であり，次元削減によるアプローチが適切ではない制約下での状態削減問題に取り組むために，観測値の生成過程を利用した環境推論によるアプローチを提案する．状態削減問題の近似解法として，強化学習でしばしば用いられる Partially Observable Markov Decision Process (POMDP) と本論文で新たに提唱する Redundantly Observable Markov Decision Process (ROMDP) の位置付けについて整理した．両者の違いは状態と観測値の生成過程にある．ROMDP は観測値から状態が生成されると仮定するため，観測値は状態を決定するのに十分な情報を持っており，報酬予測とは無関係な情報が含まれることを許容する．一方で，POMDP では状態から観測値が生成されると仮定し，現在の観測のみでは状態を決定できない．これま

で研究されてきた POMDP を仮定したエージェントは ROMDP に従う強化学習環境でコアを推定できないため，ROMDP を仮定してコアを推定するエージェントの提案を行う．

継続的な行動制御が求められる環境において，環境に ROMDP を仮定した提案エージェントと POMDP を仮定した既存エージェントについて整理した．提案エージェントと既存エージェントはベイズモデルで定式化されるため，事後分布を計算することで環境ダイナミクスの学習則が導出される．しかし，これらのエージェントモデルの事後分布を解析的に計算するのは困難であるため，自由エネルギーを最小化することで事後分布を近似する変分 EM アルゴリズムによって環境ダイナミクスの学習則を導出した．既存エージェントは全ての観測が持つ情報を予測可能なように自由エネルギーを最小化するため，推定される状態は必ずしもコアに近づかないことを理論的に確認した．一方，提案エージェントは報酬のみを予測するように観測値を状態へクラスタリングするように自由エネルギーを最小化するため，コアに近い状態で環境ダイナミクスが学習できることを理論的に確認した．そして，エージェントモデルで学習している環境ダイナミクスから計算できる将来の総報酬の分布に基づいて，探索的な行動と報酬獲得をする活用的な行動を制御する手法を提案した．将来の総報酬の分布は学習している環境ダイナミクスの精度に依存して分散が増減する．この精度が高い場合には総報酬の分布の分散は小さくなり，精度が低い場合にはこの分散は大きくなる特徴がある．この特徴を利用して，提案する行動制御法は学習している環境ダイナミクスの精度に依存して探索と活用を制御する．

5 章　数値実験

5 章は [2] の数値実験の内容であり，環境推論によるアプローチで提案したエージェントに対して，状態削減が与える戦略学習への影響を検証した．結果として，提案したエージェントはコアに近い状態を推定でき，状態を削減しない既存エージェントよりも戦略の学習速度が速いことが示された．また，コア状態のダイナミクスの変化に対しても，提案したエージェントはその変化を認識して戦略を切り替えて追従できることが示された．そして，最適な戦略とランダムな戦略の区別が付きづらくなるコア遷移の不確定性が増加する環境でも，報酬獲得の期待損失を抑えるような頑健な戦略学習ができることが示された．

これらの結果は 2 章で関連手法を整理した基準である 1) 遷移則や報酬則の非定常な環境での戦略学習と，2) 目的に基づいた状態削減を達成するものであり，提案したエージェントがこれら 2 つの基準を満たしていることを示している．また，コアに近い少数の状態で環境ダイナミクスを推定するため，目的を達成するのに最小な状態集合で遷移則を表現することができる．このことは，3) 環境ダイナミクスの推定に基づいた説明可能性は，状態削減ができない既存エージェントよりも向上していることを示している．

6 章　考察とまとめ

次元削減と環境推定によるアプローチで状態削減問題に取り組むエージェントの特徴を整理し，今後の展望について述べた．2 つのアプローチは，異なる制約で状態削減問題に取り組むため，次元削減による前処理の後に ROMDP による状態削減を組み合わせて応用することができる．このことは，実世界での応用時に課題となる観測値のスケーラビリティを解消することにつながると考えられる．また，報酬の情報が冗長，かつコアの情報が部分的にしか観測できないような POMDP と ROMDP の性質を兼ね備えた状態削減問題へ，提案したエージェントを拡張することが強化学習のさらなる発展に寄与すると考えられる．

査読つき論文

[1]. Takahashi, Kazuki, and Takashi Takekawa. ”Discriminant laplacian eigenmaps by the approximation of discriminant analysis using similarity.” Nonlinear Theory and Its Applications, IEICE 13.2 (2022): 300-305.

[2]. Takahashi, Kazuki, et al. ”Goal-oriented inference of environment from redundant observations.” Neural Networks 174 (2024): 106246.

目次