審 査 の 結 果 の 要 旨

氏　　名　　青柳　祐基

修士（工学）青柳祐基提出の論文は，「強化学習への事前知識の組み込みとシステム制御への応用」と題し，6 章から構成されている．

近年の機械学習分野の研究開発の進歩と社会実装の進展は目覚ましく，現在は第 3 次 AI ブームの最中にあるといわれている．特に強化学習については，一部のタスクにおいて人間のエキスパートを超える性能を示して以来，日進月歩で研究が進んでいる．強化学習とは，事前知識を用いずに，状態，行動および報酬の履歴のみから最適 Bellman 方程式を on-line で解くアルゴリズムの総称である．強化学習は高い自律性を持ち，事前知識を必要としないが故の汎用性を長所とする反面，探索の初期に不安定な振る舞いをし，また学習に時間を要するといった欠点により，制御問題への応用が進んでいるとは言い難い状況にある．以上を踏まえ，本論文の目的は，事前知識と強化学習の性能について検討した上で，実際に事前知識を組み込んだ強化学習の手法を提案することである．さらに，耐故障制御を想定した宇宙機および航空機の姿勢制御問題に対して提案手法を適用し，その性能を評価する．

第 1 章は序論であり，先行研究ならびに研究課題をまとめたうえで，本論文の位置づけを述べている．

第 2 章では，最適制御問題の数値解法における強化学習の位置付けを定義した上で，強化学習に関する基礎的な事項である Markov 決定過程および Bellman 方程式とその諸性質，Q-learning 等の古典的な手法についてまとめている．また，深層強化学習や model-based 強化学習，安全な強化学習等の，強化学習の発展的な概念に関する近年の研究についても言及している．

第 3 章では，事前知識とアルゴリズムの性能に関する先行研究である探索における No Free Lunch 定理とその解釈，さらには定理の拡張に関する関連研究について述べている．

第 4 章では，探索における No Free Lunch 定理の一つの拡張として，強化学習における No Free Lunch 定理を提案し，強化学習の性能を向上するには何らかの事前知識を組み込むことにより，問題領域を限定する他ないと結論づけた．

これは本論文の目的の一つである強化学習への事前知識の組み込みを支持する結果である．また，強化学習への事前知識の組み込み手法として，第１に「環境に関する知識の組み込み」として，対象の近似モデルが既知かつ報酬が二次形式で表されるという前提，すなわち環境についての不完全な事前知識が利用可能な場合に，Riccati 方程式の解を利用した状態行動価値関数の初期化による学習効率の改善手法を提案している．第２に「ドメイン知識の利用」として，制御問題の多くにみられる状態遷移が決定論的であるという前提そのものを一種のドメイン知識であると見なし，model-based 強化学習の代表的な枠組みである Dyna と，深層強化学習アルゴリズムである DQN および DDPG を組み合わせた手法を提案している．

第 5 章では，数値計算例として，環境に関する知識の組み込みについては Q-learning による double integrator の制御，ドメイン知識の利用については DQN および DDPG による double integrator の制御ならびに，より応用的な例として，DDPG による宇宙機および航空機の姿勢制御を取り上げ，提案手法の有効性を検討している．その結果，環境に関する知識の組み込みについては，学習初期の不安定な振る舞いを改善し，安全な強化学習の１つの実装例としての効果を実証している．ドメイン知識の利用については，まず，double integrator に対する計算例で，1 ステップあたりの計算時間は増えるが，学習に要する episode 数が大幅に減少する，すなわちサンプル効率が向上することを示している．さらに，より応用的な例として，宇宙機の 3 軸姿勢制御則の自律獲得を試みた結果，既存手法よりも安定かつ高効率で学習を行えることが示され，非 episode 型，すなわち online での 3 軸姿勢制御則の獲得にも成功した．また，航空機の耐故障制御を想定し，推力のみによる縦系の姿勢制御則の自律獲得についても，提案手法は，既存の手法に比べて静定時間およびピッチ角変動の最大値が約半分になるという高い性能を示すことが確認された．

第 6 章は結論であり，本論文のまとめと今後の展望について述べている．

以上要するに，本論文は，事前知識と強化学習の性能について議論した上で，実際に事前知識を組み込んだ強化学習の手法を提案し，その性能について検討している．特にドメイン知識の利用と称した決定論的環境を前提とすることで，従来の model-based 強化学習よりも高速であり，かつ制御周期の中で時間が許す限り planning を行う本論文の提案手法は，宇宙機や航空機の耐故障制御のみならず，一般の実問題への適用の可能性を示すものであり，航空宇宙工学上貢献するところが大きい．

よって本論文は博士（工学）の学位請求論文として合格と認められる．