

論文の内容の要旨

論文題目	獲得免疫系に基づいた強化学習による制御器設計に関する研究
学 位 申 請 者	細川 崇

生産工程などあらかじめ作業内容や環境が固定された状況で用いられる産業用ロボットに対し、最近では人間の代わりに日常環境で用いられる家の中の掃除を行う家庭用ロボットや、介護用ロボット、警備を行うロボットなどが数多く登場している。産業用ロボットなどでは目標や動作環境が固定されているので、通常の最適制御などによりある種の最適な行動を設定することができる。しかし、今後導入が見込まれる家庭用のロボットは運用先によって目標とする状態や目標達成に必要な政策（行動セット）が異なるため、それぞれの運用先に合わせた適切な政策を設定しなければならないが、われわれが多種多様なロボットに対して、また考えうる環境条件すべてを考慮して適切な政策を設定するのは大きな負荷となる。

本研究では、ロボットの制御器（コントローラ）の容易な構築を実現するために強化学習による手法を取り扱う。強化学習はロボットの内部状態や詳細な環境情報を与えなくとも、ロボット自身による試行錯誤の結果より自動的に適切なコントローラを学習することが可能である。一般的に目標達成に最適な政策を得るために膨大な学習時間を必要とするため、特にロボットのコントローラへの応用では最適な政策を得ることよりも学習時間の短縮が重要となる。しかし、強化学習では“次元の呪い”と呼ばれる環境認識に関する問題や、報酬や内部パラメータの初期値によっては学習がなかなか進まない、といった問題がある。

一方、生物の持つ生態機構や進化の仕組みなどを工学モデル化し、最適解探索や学習などの分野に応用する試みが盛んに行われている。その一つに免疫機構の振る舞いに着目し、

その働きをモデル化した免疫型強化学習がある。免疫型強化学習法は従来の強化学習法と比べ、特定環境において準最適解を高速な学習収束速度で得ることができる。しかし、免疫型強化学習は動作環境が連続値で表現される場合では従来の強化学習法と同じく次元の呪いによる影響を受けてしまう。これは免疫型強化学習法のアルゴリズムにおいて環境情報を離散値へ変換する必要があるためである。この変換方式として動作環境の連続値表現を一定の間隔で区切ることによって離散値表現に置き換えを行うタイルコーディングが多く用いられている。この際、状態を区切る間隔によって学習の収束速度および得られる解の質のトレードオフが発生するが、多くの場合において事前に適切な間隔を知ることはできない上、学習途中で離散化の間隔を変更することもできない。このため、事前に適切な離散化間隔を設定する必要のあるタイルコーディングによらない状態表現方法が必要となる。

さらに、制御工学で重要な安定状態を維持するといった課題においても十分な解を得ることができない。免疫型強化学習や Profit Sharing をはじめとした一部の強化学習法では、タスクの達成のための最適解を得るのではなく、実用的な解を短時間で得ることを目標に主眼をおいてアルゴリズムが構築されているからである。またその制約条件として、報酬は正の値を使用しなければならないことあげられる。安定化制御問題では報酬を与える明確なタイミングとして安定状態から不安定状態へ遷移したときが考えられる。この場合においては望ましくない状態へ遷移したため罰報酬を与える必要があるが、これまでの手法では正しく罰を取り扱うことができない。このため、安定化制御を考慮した報酬の処理法が必要となる。

本研究では、これらの問題を解決する手法を提案し、実ロボットへ適用できる学習によるコントローラの構築法を確立することが目的である。まず連続値環境を前提とした免疫型強化学習法の拡張方法を提案している。拡張したアルゴリズムが従来の離散型免疫型強化学習法の更新方式と等価であることを示し、さらに連続値環境に用いる際に利点となる状態の取り扱い方法について述べている。この提案手法を倒立振子の振り上げ制御などに適用し、従来の代表的な強化学習法と比較を行い、その有効性を示している。次に、従来の報酬割り当て関数が安定化制御問題へ適用できないことを示し、安定化制御問題へ適用する際の条件の検討を行っている。得られた条件から Profit Sharing および免疫型強化学習において有効な報酬割り当て関数の一例を提案している。提案する報酬関数を用いて倒立振子の安定化制御および RoboCup サッカーシミュレーションリーグのサブ問題である Keepaway シミュレーションに適用し、その有効性を示している。

論文審査の結果の要旨

学位申請者氏名 細川嵩
 審査委員主査 樋口幸治
 委員 ※中野和司
 委員 新誠一
 委員 桐本哲朗
 委員 内田雅文

第1章は研究の背景と目的についてである。ロボットを動作させるためには制御器設計が重要であるが、多種多様化するロボットに対して従来の制御器設計手法では対応できなくなりつつある。このためロボットの知能化技術が注目されており、その応用としてロボットの試行錯誤の結果から目標達成のための政策（行動セット）の学習を行う制御器の自動設計技術がある。その実現方法として強化学習があるが、実ロボットへの適用には学習時間の短縮化などの問題があげられている。伊藤らによる免疫型強化学習はこれらの問題を解決する手法として生物が備えている免疫機構を参考にした手法であり、Q学習やProfit Sharingなどと比べ多くの利点を備えている。しかし、連続値環境や安定化制御などへの適用については問題が残されている。これらの問題点を解決することにより、幅広い分野において学習による制御器設計手法を適用できるようにすることが本研究の目的である。

第2章は研究の基礎技術となる免疫型強化学習とその特徴についてである。モデル化を行った生物の免疫機構について獲得免疫系の働きおよび学習・記憶作用について解説し、モデル化方法およびアルゴリズムについて説明を行った。免疫型強化学習器はProfit Sharingと同等の報酬関数を用いることにより学習する政策の質が同等かつ初期パラメータに依存しない優れた手法であることを示した。また、免疫型強化学習器で学習される報酬値の傾向から相性のよい行動選択手法がルーレット選択であることを示した。

第3章は連続値によって環境情報が表現される場合についての免疫型強化学習器の改良についてである。従来の免疫型強化学習器は離散値によって状態が表現されるこ

とを前提としているが、学習を開始する前に環境に対して適切な離散化度合いを設定しないと得られる政策の質や学習収束速度に多大な影響を与える。本研究では獲得免疫系の働きを精査し、抗原の認識作用を再モデル化することにより連続状態表現をそのまま使用できるよう免疫型強化学習器の改良を行った。提案手法は学習途中でも離散化度合いを変更できる手法であり、学習結果の再利用ができる。提案手法の有効性を強化学習のベンチマーク問題である台車の山登り問題、倒立振子の振り上げ制御問題に適用して検証を行った。シミュレーション結果より、提案手法は従来の Q 学習や離散型免疫型強化学習器と比較して、学習した政策の質や学習収束速度において優位性があることを示した。

第 4 章は安定化制御問題におけるモデルフリー強化学習についての問題とその解決手法についてである。モデルフリー手法においては環境から受け取る報酬値を正の値と仮定して報酬関数などの設計が行われてきた。しかし、安定化制御問題では環境から与えられる報酬は罰報酬(負の値)であることが多く、従来の環境情報の取り扱い方や報酬関数などでは目標達成のための政策を得ることができない。本研究では安定化制御問題を考慮した状態の表現方法としてセミマルコフ決定過程を使用することと報酬関数に求められる条件および例示を行った。提案手法は免疫型強化学習器だけではなく Profit Sharing などのモデルフリー型の強化学習一般で使用できる手法である。提案手法の有効性を倒立振子の安定化制御問題、Keepaway 問題に適用をして検証を行った。提案手法はモデルフリー型の強化学習器を安定化制御問題に適用できること、および Q 学習などのモデルベース手法と比べ高速な学習収束速度を有していることを示した。

第 5 章はまとめと今後の課題についてである。本研究は多種多様化するロボットの制御器設計を簡略化のための学習制御器について取り上げ、既存手法と比べさまざまな利点がある免疫型強化学習器の幅広い応用を目的としている。従来の免疫型強化学習器は連続値環境や安定化制御への適用に問題が残されており、これらについての解決手法を提案した。

なお、上記の提案手法の有効性と実現可能性は、シミュレーションによって示されている。今後の課題としては、さらなる学習収束速度の向上や複数台のロボットが協調して動作するマルチロボット環境についての考察が必要となる。

上記の内容を纏めた本論文は博士（工学）の学位請求論文として、十分な価値があるものと認める。