動的な階層環境における強化学習エージェントの確率知識を用いた方策改善に関する研究

氏　　名	ポッマサク ウタイ PHOMMASAK UTHAI
学位論文題目	動的な階層環境における強化学習エージェントの確率的知識を用いた方策改善に関する研究
論文審査委員	主　査　教　授　塩谷浩之 教　授　前田純治 教　授　板倉賢一

論文内容の要旨

災害現場など人が入れない場所において，ロボットの活用は急速に広がっている．コンピュータの演算速度の高速化に伴い，宇宙開発やエンターテインメント等への応用研究が盛んに行われており，実質的な業務を果たすロボットも登場している．そのなかで，ロボットが自ら環境の情報を獲得し，動作を計画し実行する自律システムへの需要が高まっており，ノイズや環境変化等の不確実性が存在する実環境でのロボットの適応的学習の実現に向けた研究が注目されている．学習主体が環境と相互作用し，情報の獲得と行動を選択する枠組みが見直されている．その中で機械学習の一つである強化学習は,簡潔なアルゴリズムと強化な数字的基礎に支えられ,さまざま応用が期待されている手法である．

強化学習エージェントは，新しい環境には方策を初期化する必要がある．それに対して，統計的アプローチとして，強化学習の枠組みに状態と行動の確率モデルの導入が提案されている．具体的には，過去に学習した環境の観測データから生成される環境知識の統計モデルを構成し，知識の混合モデルから環境変化の認識と環境変化後の方策の指針を与える手法である．しかしながら，未知環境に加えて環境変化への対応には難があり，実用的な強化学習手法となってはいない．例えば複雑な実環境では，入力状態や出力行動の数を含む実験パラメータの設定が困難となり，エージェントが方策をうまく学習できない場合がある．さらには，学習システムの複雑化に伴い計算量は増大し，計算資源に制限のある実環境を想定した場合には適用できなくなるので，計算量の抑制は重要な問題となる．本研究においては，計算量を抑えながら方策改善性能を保ち，動的な環境階層型環境に適応可能な効率的な

強化学習システムの提案を通じて，未知環境への対応と実用性を備えた強化学習手法を構築し，ロボット制御におけるアルゴリズム分野に貢献することを目的としている．

本論文においては，強化学習エージェントの入力状態を決定する観測方向と出力行動の方向を増やしながらも，利益共有法を基礎にして，複雑な実環境に適応できるように報酬与え方や重みの更新式などの改良を加えた手法を提案している．その一つとして，階層型環境適応用のエピソードの二次元化するなどのパラメータの新たな更新法の導入をする．加えて，エージェントの観測データからなる同時分布を構成要素として混合分布を用いて，未知環境に適応できるように方策改善を行うための手法を構築する．さらに，方策改善性能を維持しながらも計算量を抑制するために，混合分布の構成要素を統計的手法に基づいて少ない数で選択できる分布クラスタリングを新たに導入する．これらの提案や導入によって，既存知識から適応的に選択された環境知識による混合モデルを用いて，階層化された 3 次元環境においてエージェントが効果的に適応できることが示された．本研究において導入した統計的アプローチによって，階層構造で未知な環境に対応でき，かつ実用性のあるアルゴリズムを備えた強化学習手法が実現した．

ABSTRACT

With the increasing use of rescue robots in disasters, such as earthquakes and tsunami, there is an urgent need to develop robotics software that can learn and adapt to any environment. Reinforcement Learning (RL) is often used in the development of robotic software. RL is a field of machine learning within the computer science domain; moreover, many RL methods have been proposed recently and applied to a variety of problems, where agents learn policies to maximize the total number of rewards determined according to specific rules. In the process whereby agents obtain rewards, data consisting of state-action pairs are generated. The agents' policies are improved effectively by a supervised learning mechanism using a sequential expression of the stored data series and rewards.

Typically, RL agents must initialize policies when they are placed in a new environment, and the learning process starts afresh each time. Effective adjustment to an unknown environment becomes possible using statistical methods, such as a Bayesian network model, mixture probability, and clustering distribution, which consist of observational data

for multiple environments that the agents have learned. However, adapting to environmental change, such as unknown environments, is challenging. For example, setting appropriate experimental parameters, including the number of the input status and the output action, becomes difficult in complicated real environments, and that makes it difficult for an agent to learn a policy. Furthermore, the use of a mixture of Bayesian network models increases the system's calculation time. In addition, due to limited processing resources, it becomes necessary to control computational complexity.

The goal of this research is to create an efficient and practical RL system that is adaptive to unknown and complex environments, such as dynamic movement environments and multi-layer environments. In addition, the proposed method attempts to control computation complexity while retaining system performance.

In this study, a modified profit-sharing method with new parameters, such as changing reward value, is proposed. A weight update system and changing the dimension of the episode data make it possible to work in dynamically moving multi-layer environments. A mixture probability consisting of the integration of observational environmental data that an agent has learned within an RL framework is introduced. This provides initial knowledge to the agent and enables efficient adjustment to a changing environment. A clustering method that enables selection of fewer elements has also been implemented. This reduces computational complexity significantly while retaining system performance. By statistical-model approach, an RL system with a utility algorithm that can adapt to unknown multi-layer environments is realized.

論文審査結果の要旨

コンピュータの演算速度の高速化に伴い，工場生産だけではなく災害における救出作業や介護の現場まで，ロボットの活用は急速に広がりを見せている。その中でロボットが自ら環境の情報を獲得し，動作を計画し実行する自律システムへの需要が高まっており，ノイズや環境変化等の不確実性が存在する実環境における適応的学習の実現に向けた研究が注目されている。学習主体が環境と相互作用し，情報の獲得と行動を選択する枠組みが見直されているなかで，環境のマルコフ性に対応した数字的基礎に支えられた強化学習に対して，様々な基礎および応用研究が行われている。

強化学習における学習主体はエージェントと呼ばれ，新しい環境に対し方策を初期化して学習をやり直す必要がある。それへの対応策として，強化学習の枠組みに状態と行動の確率モデルの導入が提案されている。具体的には，過去に学習した環境の観測データから生成される環境知識の統計モデルを構成し，知識の混合モデルから環境変化の認識と環境変化後の方策の指針を与える手法である。しかしながら，未知環境に加えて環境変化への迅速な対応には適応せず，計算量の問題からも実用的な強化学習手法となってはいない。

本提出論文においては，計算量を抑えながら方策改善性能を保ち，動的な環境階層型環境に適応可能な効率的な強化学習システムの提案を通じて，未知環境への対応と実用性を備えた強化学習手法を構築し，ロボット制御におけるアルゴリズム分野に貢献することを目的としている。強化学習エージェントの入力状態を決定する観測方向と出力行動の方向を増やしながらも，利益共有法を基礎にして，複雑な実環境に適応できるように報酬与え方や重みの更新式などの改良を加えた新たな手法を提案している。その中では，階層型環境適応用のエピソードの二次元化するなどのパラメータの新たな更新法を導入している。加えて，エージェントの観測データからなる同時分布を構成要素とした混合分布を用いて，未知環境に適応するための方策改善法を構築している。さらに，方策改善性能を維持しながらも計算量を抑制するために，混合分布の構成要素数について，統計的手法に基づいた分布クラスタリングを新たに導入している。これらの提案や手法導入により，既存知識から適応的に選択された環境知識による混合モデルを用いて，階層化された 3 次元環境においてエージェントが効果的に適応することを示している。本研究における提案を含めた統計的アプローチによって，階層構造で未知な環境に対応し，実用性のあるアルゴリズムを備えた強化学習手法を実現している。よって本研究は，情報工学におけるソフトコンピューティング分野に大きく貢献をしており，本提出論文は，学位論文として認められる。