異なる環境に適用可能な知識を考慮した強化学習手法の構築

篠塚　敬介

近年，情報処理技術の発達により，コンピュータプログラムによって高度な機械制御が可能になってきている．この状況を背景に，あらかじめ決められた動作をするのではなく，状況に応じた動作をするような複雑な処理をロボットに行わせることへの要求が増加している．そこで，状態ごとの行動決定ルールをプログラムする代わりに，目的状態を設定することによって，目的状態に到達するための行動ルールを自動的に発見させる学習手法である強化学習が注目されている．

しかし, Q 学習などの基本的な強化学習手法では, 以前の環境での知識をなにも記憶せず, 新たな環境でもはじめから学習してしまい，学習効率が悪いという問題がある．この問題を解決するため，状態集合がほぼ一致している環境で有用となる知識の再利用を行う手法が提案されている．しかし，こうした手法では，環境が大きく変化すると使えなくなってしまう．そこで，本研究では，ドメイン（問題領域）が同じであれば，環境において不変な知識を再利用することで，環境の大きな変化に左右されずに学習効率を向上させることのできる新たな強化学習手法を提案する．本論文では，この手法をドメイン学習と呼ぶ．

ドメイン学習は，ドメインが同じ環境における状態の変化量と行動の関係を不変な知識であると仮定し，これを D 値として Q 学習のアルゴリズムにおける行動の選択に反映することで実現した．行動の選択には，D 値を最重視した局所的な見方による行動の選択と，再訪率を考慮した大域的な見方による行動の選択の 2 つがある．

学習効率の向上を検証するため，状態集合の一致しない，それぞれ異なるレーストラックを環境として，記憶をもたない Q 学習と記憶をもつドメイン手法の学習のはやさを比較する実験を行った．実験の結果，ドメイン学習はより効率の良い学習を行えることがわかった．また，ドメイン学習で D 値を最重視した行動を選択した場合，再訪率を考慮した行動を選択した場合に比べて効率が良かったが，失敗することがあった．以上のことから，D 値を最重視したドメイン学習と再訪率を考慮したドメイン学習を並列させることで，新たな環境において最も効率の良い学習を行うことができると考えられる．

今後の課題・展望として，適応的な再訪率の設定，D 値を最重視した行動を選択するドメイン学習で発生する失敗についての詳しい調査・実験，部分観測マルコフ決定過程として定義された環境でのドメイン学習の適用などが挙げられる．

（指導教員　若林啓）