平成 24 年 01 月 10 日

情報・知能工学専攻	学籍番号	083717
申請者氏名	小泉 勇人	

指導教員氏名	村越 一支

論 文 要 旨 (修士)

論文題目	動的な新状態生成・削除と時間経過を考慮した状態選択による効率的な強化学習手法

時間経過によって同じ行動をしても報酬が得られたり，得られなかったりすることがある．このような例としては，インターネットサイトへのアクセスにおいて，混雑緩和のため一定時間以上経過してからでないと再びアクセスすることができないという制限がある場合や，人間や動物の行動を研究する分野で用いられる固定間隔 (FI: fixed interval) スケジュールや変動間隔スケジュール (VI: variable interval) が挙げられる．こうした問題では，報酬が得られる時間と得られない時間を正確に捉え，報酬が得られる時間になったときに，その報酬につながる行動ができれば効率よく報酬を得ることができる．本研究では強化学習を用い，上記のような性質を持つ問題において効率的な行動を学習する手法を検討する．

時間経過を考慮して課題を解決しようとしている例としては，KDDI(2011) が開発した無線端末における通信技術が挙げられる．これは過去の通信履歴から回線が混雑する時間帯を判断し，通信時間を決定するという技術である．しかし，この技術はいつかダウンロードが完了すればよいという程度の不急のアプリケーションについて考えられており，本研究で検討する時間的な効率性という点については考慮されていない．

前に報酬を得てから一定時間以上経過してからでないと再び報酬が得られないという制限がある場合，同じ行動をしても時間経過によって結果が異なる．つまり，報酬が得られるときと得られないときとでエージェントは異なる状況に置かれていると解釈できる．

したがって，このような性質を持つ課題を効率よく学習するためには，新状態を生成し，学習を行っていく必要があると考えられる．また，状態の生成については，人間があらかじめ問題を分析し，それに応じて新状態を生成するのは非効率であるため，動的に行うことが望まれる．しかしながら，予備実験として，既存の強化学習手法の一つである Actor-critic を用いて，置かれている状況によらず状態を変えることなく学習を行ったところ，効率的な行動選択をしていないことが分かった．

そこで，本研究では一定時間経過後に再び報酬が得られる環境において，動的な新状態生成・削除と時間経過を考慮した状態選択を行うことで，状況に応じた行動を学習する効率的な強化学習手法を提案する．動的な状態生成を行うための指標として，本研究では強化学習内で用いられる TD 誤差 (Temporal Difference Error) を利用することで事前にデータを収集せずに新たな状態の生成を行う．また，新しく生成した状態と元の状態で差が見られず，不要と判断できる場合には削除を行う．更に，時間経過によって置かれる状況が異なることから，これを考慮した状態選択を行う．具体的には報酬が得られる時間間隔を学習し，その時間間隔および学習進行度に応じて状態の選択確率を決定する手法を提案する．

提案手法の有効性を示すため従来手法との比較実験を行った結果，一定時間以上経過してから報酬が得られる問題においては従来手法と比較して大幅に合計報酬値が上回った．また，本研究で検討する問題のような性質を持たない問題においても対応できるかを検証するため，Murakoshi and Mizuno(2004) が用いた迷路問題およびアームロボット問題を用いて比較を行った．その結果，従来手法と同程度の結果が得られることを確認した．