平成 18 年 1 月 12 日

知識情報工学専攻	学籍番号	991014
申請者氏名	稲生 亮二	

指導教員氏名	村越 一支

論 文 要 旨 (修士)

論文題目	パラメータ更新ステップ幅の自動制御による強化学習の環境への順応

強化学習とは、与えられた環境のなか、エージェントが報酬に繋がる行動を試行錯誤を通じて学習し、最終的に最適となるような方策を決定するアルゴリズムである。この研究分野において、環境変化への追従性を向上させるための手法が幾つか提案されている。これらの手法では、固定ステップ経過毎 (以下、パラメータ更新ステップ幅と呼ぶ) に、報酬の減り具合を用いて強化学習パラメータを更新する。しかし、環境の変化に合わせてパラメータ更新ステップ幅も修正すれば、環境変化後の行動改善をより素早く行なえるだろう。実際に、パラメータ更新ステップ幅が違うことで獲得される報酬の合計にどのような違いが生じるか予備実験を行なった。その結果、環境変化直後ではパラメータ更新ステップ幅が短い場合に得られる報酬が多くなった。更に、報酬が繰返し周期的に得られる場合、パラメータ更新ステップ幅がその周期の倍数の時に、最終的に得られる報酬の合計が多くなることも解った。これは、パラメータ更新ステップ幅を報酬が得られるまでの最短ステップ数に合わせることで、合計報酬のずれが無くなったためだと考えられる。報酬を元に強化学習パラメータを更新するため、このような報酬のずれがあるとパラメータ更新の際に弊害となる。

本研究では、強化学習パラメータを固定ステップ毎に更新するような強化学習アルゴリズムに対し、与えられた環境に順応するような強化学習アルゴリズムの実現を目的とする。まず、学習の初期段階では、環境全体が学習できていないため、行動はランダムである。この時にパラメータ更新ステップ幅を制御しても意味が無いため、学習の初期段階では学習が安定するまでパラメータ更新ステップ幅は制御しない。学習が安定した後に環境変化が起こった場合、パラメータ更新ステップ幅を最短にして早急な再学習を行なう。以降、学習が再び安定するまではパラメータ更新ステップ幅を短めに設定し、学習が再び安定したらパラメータ更新ステップ幅を長めに設定する。最終的なパラメータ更新ステップ幅には獲得される報酬の周期の倍数を設定する。報酬の周期は、1 ステップ毎に得られる報酬情報の時系列データをフーリエ変換することで取得できる。なお、環境変化の検知については、既存アルゴリズムと同様に、パラメータ更新ステップ幅毎に計測される報酬の減り具合を用いて表わす。

本手法を Schweighofer & Doya(2003)、および Murakoshi & Mizuno(2004) の提案する二つのアルゴリズムに対し適用し、計算機実験を行なった。取り扱う問題設定として、報酬獲得の周期性が容易に認められる迷路問題を幾つか用意した。従来のパラメータ更新ステップ幅固定の場合の実験結果と比較したところ、環境変化後の学習において、提案手法を適用した場合に多くの報酬が得られた。以上より、環境に応じパラメータ更新ステップ幅を自動制御することは、より多くの報酬を得るために有効であるといえる。