平成28年度 情報工学コース卒業研究報告要旨

金森 研究室	氏　　名	西尾　宣紀
卒業研究題目	統計的学習に基づく強化学習アルゴリズムの考察	

強化学習とは機械学習の枠組みの一つであり、教師あり学習、教師なし学習のどちらにも属さない学習手法である。一般的な教師あり学習が教師信号を目印に学習を進めていくのに対して、強化学習ではエージェントと呼ばれる主体が行動を行った結果得られた報酬を目印にして学習を進めていくものである。強化学習の学習フェーズは大きく方策推定と方策評価の二つの部分に分けることが出来る。方策評価の部分ではエージェントが実際に行動をとり、得た報酬を用いてそのエージェントの行動方策がどの程度有用であるかを状態や行動の価値という形で検証する。一方、方策推定のフェーズではその価値に基づいて行動方策を得る。この枠組みの大きな特徴として、環境モデルが未知でも学習が可能であることがあげられる。現実の制御問題において、環境モデルは未知であることが多いため、AlphaGo に代表されるゲーム AI やロボット制御の分野において応用が期待されており、強化学習をより効率よく行うためのアルゴリズムが現在多数提案されている。

本研究では、提案されている強化学習アルゴリズムのうち代表的なものを統計的学習の観点から評価し、それを踏まえた上で最適な学習アルゴリズムについて考察する。まずはじめに一般的な強化学習の枠組みについて概観し、最も広く利用されている「Temporal Difference Learning(TD 学習)」の学習規則の導出を行う。次に状態の価値関数を線形近似する状況を考えた上で、代表的な強化学習アルゴリズムに関して、統計的学習の観点から意味付けを行う。続いてセミパラメトリックな統計的学習に基づいて強化学習を定式化し、それをもとに最適な強化学習アルゴリズムについて考察を行う。この時簡単のために、状態価値関数を線形近似した時の近似誤差は無いものとし、またエージェントの行動方策は固定し、方策評価のフェーズに関してのみ考えるものとする。

まず数値実験により、導出したアルゴリズムの性能評価を行った。問題設定として、5つの鎖状に連なる状態をエージェントがランダムに動き、それぞれの状態に到達した時に報酬を得る状況を考えた。比較したアルゴリズムは、TD Learning と TD Learning の高速化を図った Accelerated TD Learning、求める推定量に関する漸近推定分散を最小にする Optimal TD Learning の三つである。この問題設定において状態価値関数をどれだけ正確に推定できるかでアルゴリズムの性能を比較した結果、Optimal TD Learning が他の二つのアルゴリズムより良い性能を示すことが確認できた。

Optimal TD Learning が推定量の漸近推定分散を最小化するのは、状態価値関数を近似した時の誤差が無く、尚且つ方策を固定した時のみである。そこで次に、方策をボルツマン選択で変化させた時の各アルゴリズムの挙動を数値実験で確認した。問題設定は先のものと同じで方策のみ変化させたものと、先の設定の状態数を 10 に増やしかつ方策を変化させたものの 2 つを用いて実験を行った。1000 回の状態遷移で得られた報酬の合計で性能を評価した結果、どちらの問題設定でも Accelerated TD Learning が最も性能が良く、次に Optimal TD Learning が良いという結果が得られた。

Optimal TD Learning は、漸近分散を最少にするための仮定が崩れた時に、収束性や収束レートがどのようになるのかまだわかっていない。今後の課題として、これらを理論的に求めることが挙げられる。