帰納された知識をガイドに用いた強化学習の研究

指導教員 助教授　犬塚　信博
システムマネジメント工学科 UE コース
13119639
韓　相斌

1．はじめに

未知の環境内で行動するエージェントが環境を同定するための学習手法として強化学習がある．エージェントは状態から行動へのマッピング―政策を学習し，最適な政策を見つけることを目的とする．本研究では，エージェントがある環境で学習した結果（最適な政策）から一般的な知識を獲得し，別の環境でその知識を利用する方法を提案する．また，その有効性を実験により確認する．

2．属性つきマルコフ決定過程

状態と行動を持ち，マルコフ性を満たす（直前の状態と行動のみで，次状態への遷移確率が決まる）強化学習タスクはマルコフ決定過程（MDP）と呼ばれる．ある MDP で得た政策は別の MDP に利用することができない．しかし，状態が属性を持ち，これを観測できるならば，この属性を共有する別環境に政策を適用できる可能性がある．本研究ではこれを仮定した環境を属性つき MDP と呼ぶ．

環境の状態を属性を用いて記述することによって，一般的な知識を獲得し，それを行動のガイドとしてエージェントに与えることで，エージェントの学習効率を向上させることが本研究の目的である．そのため，属性を共有した複数の MDP を形式化し，いくつかの MDP で得た政策から知識を得る方法と，得られた知識を別環境で利用する方法を提案する．

3．強化学習の結果からの知識獲得

エージェントは各状態で行動することよって報酬を得る．この報酬を各状態と行動の価値として累積することで学習する．Q 学習では，各状態でその後将来にわたって，得られる報酬の期待値を Q 値と呼び，これを学習する．

属性つき MDP を満たす環境の状態は各属性の属性値（これを属性ベクトルという）によって記述できる．この属性ベクトルとその状態における最適な行動（最適な政策の値である）の対を一つの事例とする．環境の全ての状態に関する事例からなる事例集合を使って，属性から最適な行動を予測する知識を求めることができる．

4．提案する知識の利用法

次に，得られた知識を用いる方法を提案する．

（1）．知識に依存した行動選択法(知識ガイド法)．これは各状態において，知識が予測する行動と Q 値が最大となる行動を確率的に選択させる政策である．環境の全ての状態を把握するため，知識だけでなく，ある程度の確率でランダム的な行動も取らせる．

（2）．知識を用いた Q 値初期化法(知識初期化法)．強化学習アルゴリズムの最初の段階では，価値の初期化を行う．そこで，この提案法では任意に初期化するのではなく，各状態で知識が予測する最適な行動との対に高い Q 値を与えて，初期化する．

5．実験と結果

実験は初期状態を固定しないハノイの塔の環境を用いた．この環境は属性つき MDP モデルとして定式化できる．実験で利用した知識はハノイの塔の円盤が 4，5 枚の環境で行った強化学習の結果から帰納したもので，円盤 6 枚の環境での最適な行動の分類精度が 63.46%であった．この環境で二つの提案手法，そして，従来の強化学習の手法（Q-学習＋ε-greedy 法）を用いて，10 回ずつ実験した．実験結果として，それぞれの手法を使った場合の Q 値の収束様子を図１に，目標到達度を図２に示す．

6．まとめと今後の課題

知識と強化学習の組合せたことによって，

1．強化学習の結果を知識を介し別環境に利用すること
2．これによって学習の効率改善があったこと

を確認した．実験で使った知識の分類精度が６割であったが，それでも，学習の効率がよくなっている．知識初期化法は特に優れる．

今後，強化学習の効率改善のしくみ，知識の分類精度の変化からの影響について，調べる必要がある．

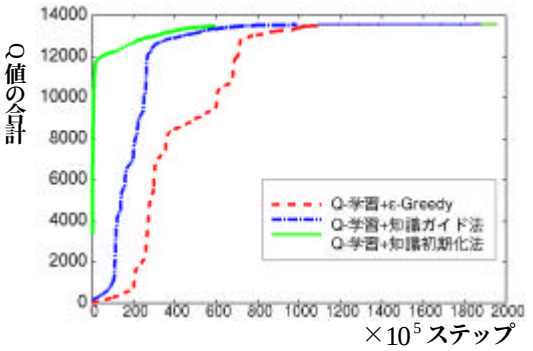

図１　円盤が６枚の場合の Q 値の収束状況
縦軸：各時点の各状態における最大と Q 値の合計（１０回平均）

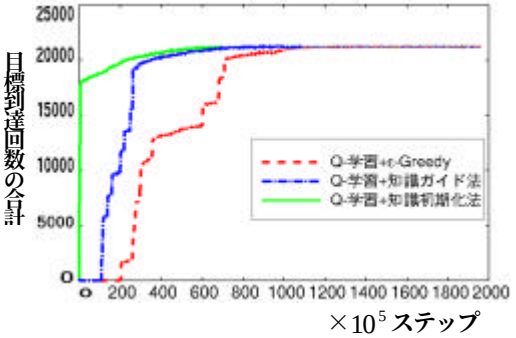

図２　円盤が６枚の場合の目標到達状況
縦軸：10^5 ステップ毎の目標達成の回数（10 回合計）