平成 17 年度　卒業研究概要

入学年度　平成 14 年	学籍番号　14117688	氏名　高木 章広
論文題目　述語論理表現を用いた強化学習結果の一般化と利用		和田・犬塚　研究室

1. 目的

強化学習とは，動物の適応的な行動獲得を工学的観点からモデル化したものである．強化学習では，環境の中で自律的に行動するエージェントが，報酬を手がかりとして，人間から教わることなくみずから行動を学ぶ．

学習された行動は，学習した環境に依存する．ある環境で学習が終わったエージェントに，新しい環境が与えられると，エージェントは新しい環境にあった行動を最初から学習しなおす必要がある．

[韓 '05] は，ある環境での学習結果を決定木の形の知識に変換し，それを新しい環境で利用することにより，新しい環境での学習を効率的に行う方法を提案した．しかしこの手法では，エージェントの行える行動の数が，環境や状態によらず一定である場合しか扱うことができなかった．

そこで本研究では，行動の数が変化する場合も扱えるような，述語論理で表現されたルールを用いる．ここでは，学習結果をルールに変換する手法を提案し，またこれに加えて，ルールを新しい環境で利用する手法を提案する．

2. 原理

2.1 強化学習

強化学習では，学習と意思決定を行うエージェントと，エージェント外部の環境というモデルを扱う．エージェントは環境の状態を観測し，それに応じて行動を実行する．すると環境の状態が変化する．エージェントの行動によってあらかじめ設定された目的が達成されたとき，エージェントに報酬を与える．エージェントは試行錯誤をくりかえし，その結果，どの状態でどの行動を行うと最も報酬を得やすいか，という知識—最適方策を獲得する．

強化学習の用語を次にまとめる．

- Q 値は，状態-行動対に対する価値である．報酬を得やすいものほど高い．各状態で最も Q 値の高い行動が，その状態に対する最適な行動である．
- Q 学習は，Q 値を獲得するための学習方法である．Q 値の推定値を何度も更新しながら，真の Q 値に近づけていく．
- ϵ-greedy 選択は，エージェントの行動選択方法である．高確率で Q 値の推定値が最も高い行動を選び，低確率でランダムに行動を選ぶ．

2.2 ILP：帰納論理プログラミング

正事例，負事例，およびそれらに関連した背景知識を ILP システムに与えると，正事例に共通してみられ，負事例にはみられないルールを発見する．正負事例や背景知識，ルールは述語論理で表記される．

3. 提案手法

3.1 学習結果からルールを生成する手法

ある環境で強化学習を行った結果，その環境に特有の最適方策が得られる．

ここで，すべての状態-行動対を述語で表記し，最適方策に含まれる状態-行動対を正事例，含まれない状態-行動対を負事例とし，それらを全環境に共通の背景知識とともに ILP システムに与えて一般化する．それによって，正事例だけにみられ，負事例にはみられないルールを生成する．ルールによって，状態-行動対が最適方策かそうでないかを判別できる．

3.2 ルールを学習に利用する手法

ルールを得た環境と類似した環境，すなわち，状態-行動対と背景知識を記述する述語が元の環境と共通している環境が与えられたとする．

観測した状態と，そのとき実行可能な各行動との対を述語で表してルールに与え，最適方策と判定された行動を実行する．ただし，新しい環境においてルールが常に正しいとは限らない．学習を進めると，ルールに頼らず行動する．ここでは 3 種類の方法を提案した．

1. Q 値を初期化する：Q 値を初期化するとき，最適方策と判定された状態-行動対に対する値を他より少し大きな値にする．行動は ϵ-greedy 選択で行う．
2. 一度だけ優先する：最適方策と判定された状態-行動対を実行する．ただし過去にその行動を行ったことがあれば，ランダムに行動を選択する．
3. 学習するまで優先する：最適方策と判定された状態-行動対を実行する．ただしその行動に対する Q 値の学習が完了していれば，ランダムに行動を選択する．

4. 実験

いくつかの積木をエージェントが移動させ，特定の積み方に直す環境を対象として実験した．以下の手順で実験を行い，ルールを利用する 3 種類の方法で，それぞれ学習効率が改善されているかどうかを調べた．

1. 積木が 4 個の環境で，普通の強化学習を行う．
2. 得られた最適方策から，ルールを生成する．
3. 積木が 5 個の環境で，ルールを利用した強化学習を行う．

報酬を与える条件をさまざまに変えて実験を行ったところ，次の結果が得られた．

条件によって，学習効率が改善される場合とされない場合があった．生成されたルールの精度が低いと，効率が改善されないことが多かった．

5. まとめ

- 強化学習の結果得られた最適方策を，述語論理を用いて一般化し，最適な行動を判別するルールを生成する方法を提案した．
- 生成したルールを利用して，元の環境と類似した別の環境において，効率よく学習を行うための 3 種類の方法を提案した．
- 既存の手法と比較して，提案手法では行動の数が変化する問題を扱うことができる．
- 提案手法によって，いくつかの場合で学習効率を改善できることを確認した．生成されたルールの精度が低いとき，効率が改善されない場合があった．

生成されたルールの性質と学習効率の関連性を調べ，提案手法を改良することが今後の課題である．

参考文献

[韓 '05] 韓相斌, 帰納された知識をガイドに用いた強化学習の研究, 名古屋工業大学 平成 16 年度卒業論文, 2005.