平成 14 年度　卒業論文概要

入学年度	平成 11 年度	学生番号	11117612	氏名	石黒 誉久
卒業研究題目	コスト付きマルコフ決定過程のための強化学習アルゴリズム			和田　研究室	

はじめに

エージェントが知識を獲得するための学習手法として強化学習がある．このエージェントは状態と行動の対に対して，その時得られた報酬を用いて学習し報酬の期待値の最大化を目的とする．本研究では，エージェントの行動系列が有限長に自然に分割できる環境（エピソード型タスク）において，行動にコストが発生する場合の強化学習アルゴリズムの利用について検討する．

コスト付きマルコフ決定過程

有限な状態と行動を持ち，マルコフ性を満たす強化学習タスクは有限マルコフ決定過程（有限 MDP）と呼ばれる．また，本研究ではマルコフ性を満たし，行動にコストを伴い，最終状態と呼ぶ特殊な状態がある場合を考える（コスト付き有限 MDP と呼ぶ）．エージェントの行動系列は終端状態に達した所で自然に分割され，これをエピソードと呼ぶ．

即ち，コスト付き MDP は 6 つ組 $(S, S^+, A, \mathcal{P}, \mathcal{R}, \mathcal{C})$ である．S は有限の状態集合，S^+ は有限の終端状態，A は有限の行動集合，$\mathcal{P}$ は状態遷移確率，$\mathcal{R}$ は報酬関数，$\mathcal{C}$ はコスト関数である．

通常の MDP で，割引期待収益の最大化を目的とするのに対し，本研究ではコストが発生する環境において，総報酬から総コストを引いた利益の最大化について考える．

アルゴリズム

強化学習手法として Q-learning，R-learning，Profit Sharing に対して，コスト付き MDP を扱う方法を検討する．Q-learning は状態と行動の対からなる価値が最適価値へと収束することを保証している．R-learning は割引のない Q-learning よりも性能のよいアルゴリズムである．Profit Sharing は報酬獲得時にそれまでの系列すべてを強化するので学習伝搬が早い．

以上のアルゴリズムをコスト付き MDP 問題に適用する方法として，1) 総報酬から総コストを引いた利益をエピソードの最後で報酬として獲得すると考える方法（先送り更新），2) コストを負の報酬として扱い，そのつど更新する方法（逐一更新），3)Profit Sharing における信用割当関数を変える方法を考える．また 3) の方法として，3-A) 報酬を利益に変えて更新 (利益等比減少法)，3-B) 時間による割引でなく，コストによる割引をおこなう方法 (収益コスト比減少法)，3-C) 各状態行動の強化を本来受け取る利益によって更新する方法 (実効利益等比減少法) を考える．この内，1) を利用すると MDP ではなくなるが実験による効果をみる．Q-learning と R-learning にはの 1) と 2) のみを適用できる．Profit Sharing では 1) は 3-A) と一致する．2) は中間報酬を扱えないため適用できない．3) で挙げた 3 つの方法を適用する．

実験と結果

実験は図 1 の迷路問題を用いた．環境はコスト付きマルコフ決定過程モデルであり，状態は図 1 の各マス，s_1 のマスは開始状態を意味する．G のみが終端状態であり，行動は，上下左右の隣のマスに遷移する行動 a_{one} と上下左右の壁まで遷移する行動 a_{wall} をもつ（$a \in \{$ 上，下，左，右 $\}$）．状態遷移確率は決定的とする．報酬は終端状態でのみ 1.0 発生し，コストは a_{one} には 0.01，a_{wall} には 0.12 掛かる．

この環境で上記の手法を用いて実験した．結果は探査（学習中の振る舞い）と知識利用（学習後の振る舞い）として図 2,3 に示す．

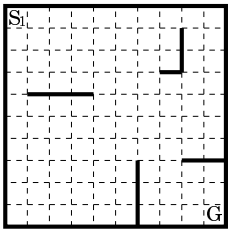


図 1．迷路問題

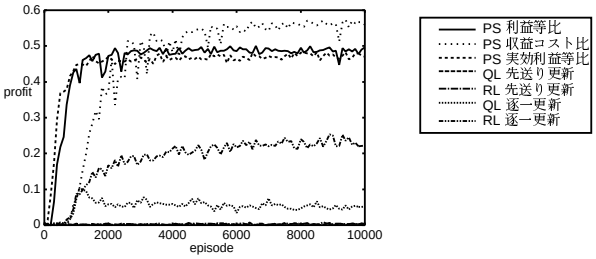
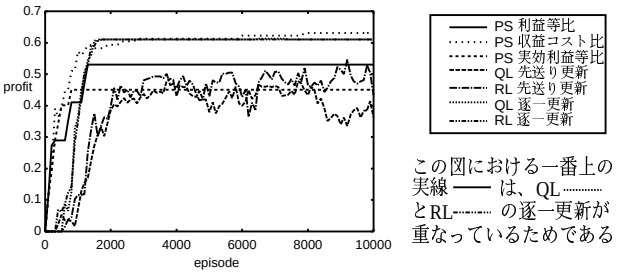


図 2．探査
パラメータ:$\gamma = 0.9, \alpha = 0.1, \beta = 0.1, t = 0.2$



この図における一番上の実線 —— は、QL ····· とRL ····· の逐一更新が重なっているためである

図 3．知識利用
パラメータ:$\gamma = 0.9, \alpha = 0.1, \beta = 0.1, t = 0.2$

Q-learning，R-learning の先送り更新は MDP でないため，実験においてもよい結果とならなかった．Q-learning，R-learning の逐一更新では知識利用においてよい結果となった．Profit Sharing は探査，知識利用ともに安定してよい結果となった．その中でも収益コスト比減少関数が最もよい性能を示した．また $\gamma = 1.0$ での Q-learning 逐一更新はをおこなったところ収束までの時間がかかってしまうが，$\gamma = 0.9$ よりも性能がよく，どの手法よりも利益を最大としていた．

まとめ

強化学習アルゴリズムをコスト付き MDP に用いる方法を研究した．

利益最大化問題に対して，学習中での性能をよくするには Profit Sharing を用いるのが適している．なかでも収益コスト比減少関数が最もよい．また，学習後に最適な行動の価値や重みを選択していく上では，学習速度を必要とする場合も Profit Sharning 収益コスト比減少関数が優れている．$\gamma = 1.0$ での Q-learning 逐一更新は最も高い利益を獲得するのには優れているが，学習に時間がかかるという欠点を持っている．ただし，状態行動価値の収束性はこの場合保証されない．