3F4-04

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

Reinforcement Learning for Markov Decision Processes with Differing Cost

石黒 誉久[*1] 松井 藤五郎[*2] 犬塚 信博[*3] 和田 幸一[*3]
Takahisa Ishiguro Tohgoroh Matsui Nobuhiro Inuzuka Koichi Wada

[*1]奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

[*2]東京理科大学理工学部経営工学科
Department of Industrial Administration, Faculty of Science and Technology,Tokyo University of Science

[*3]名古屋工業大学大学院工学研究科情報工学専攻
Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology

Reinforcement learning methods for environment including actions with differing costs are investigated.Through experiments we examined treatment of this problem with Q-learning, R-learning, and Profit Sharing. Profit Sharing with a credit assignment functions considering costs is shown to have good performance in a practical sense.

1. はじめに

強化学習 [RL 98] は，エージェントが知識を獲得するための学習手法である．エージェントは，状態と行動の対に対して，得られる報酬の期待値が最大となるように学習する．本研究では，行動にコストが発生し，かつエージェントの行動系列が有限長に自然に分解できるエピソード型タスクにおいて，総報酬から総コストを引いた利益 (profit) を最大化する問題について述べる． 従来の強化学習におけるエージェントの目標は最終的に受け取る報酬を最大化することである．**収益** R_t は報酬の系列の何か特定の関数として定義される．最も単純な場合には，この収益は報酬の合計である．

$$R_t = r_{t+1} + r_{t+2} + r_{t+3} + \cdots \tag{1}$$

また，割引率を用いる場合もある．ここで割引率は将来の報酬が現在においてどれだけの価値があるかを決定する．

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{2}$$

しかし，このような収益や割引収益を最大化すればいいという問題ばかりではない．実世界においては行動にコストがかかる場合がある．このような問題に対して本研究ではコストを扱う環境をモデル化し，総報酬から総コストを引いた利益を最大化する強化学習アルゴリズムを提案し，実験により評価する．

2. コスト付きマルコフ決定過程における強化学習問題

ここでは，従来の強化学習環境であるマルコフ決定過程をコストを扱った環境に拡張し，本研究での目標である総報酬から総コストを引いた利益を定義する．

連絡先: 石黒 誉久，奈良先端科学技術大学院大学情報科学研究科， 〒 630-0192 奈良県生駒市高山町 8916-5,Tel:0743-72-5312, taka-is@is.aist-nara.ac.jp

2.1 コスト付きマルコフ決定過程モデル

ある時刻での応答がその直前の状態と行動のみにより決定する場合，マルコフ性を持つという．強化学習の環境は，このようなマルコフ性を満たしたマルコフ決定過程 (MDP) によってモデル化される．マルコフ決定過程は 4 つ組 $(S, A, \mathcal{P}, \mathcal{R})$ であり、以下のように定義される．

- S は環境の状態の集合．
- A はエージェントが実行可能な行動の集合．
- $\mathcal{P} : S \times A \times S \to [0, 1]$ は状態遷移確率関数である．$\mathcal{P}(s, a, s') = \Pr(s_{t+1} = s' | s_t = s, a_t = a)$ は現在状態 s にあり行動 a をとったときに次状態 s' に遷移する確率である．
- $\mathcal{R} : S \times A \times S \to \Re$ は報酬関数である．$\mathcal{R}(s, a, s')$ は現在状態 s にあり行動 a をとって次状態 s' に遷移したときの報酬の期待値である．

マルコフ決定過程をコストを扱った問題に拡張する．ここで新しく定義するものはコストである．コストは各状態において行動をとったときに環境より発生するものとする．また，状態，行動，報酬の列 $s_0, a_0, r_1, \cdots, s_{T-1}, a_{T-1}, r_T, s_T$ をエピソードと呼び，T は終端ステップである．このアプローチは終端ステップに関する自然な概念が存在する場合に意味をなす．各エピソードは**終端状態**と呼ばれる特殊な状態で終わる．この終端状態に続いて，標準的な開始状態あるいは開始状態の標準的な分布のサンプルへのリセットがおこなわれる．この種のエピソードを伴うタスクは**エピソード型タスク**と呼ばれる．ここでは，エピソード型タスクで考える．

コスト付きマルコフ決定過程モデルを 6 つ組 $(S, S^+, A, \mathcal{P}, \mathcal{R}, \mathcal{C})$ で表す．ただし $S, A, \mathcal{P}, \mathcal{R}$ は上と同様とし，$S^+, \mathcal{C}$ は次のとおり定義する．

- $S^+ \subseteq S$ は環境の終端状態の有限集合
- $\mathcal{C} : S \times A \to \Re^+$ はコスト関数である．$\mathcal{C}(s, a)$ は現在状態 s にあり行動 a をとったときのコストであり，正の実数値である

2.2 利益

従来の強化学習での目標は最終的に受け取る報酬の最大化であったが，本研究での目的は最終的に受け取る利益の最大化である．状態，行動，コスト，報酬の列 $s_0, a_0, c_1, r_1, \cdots, s_{T-1}, a_{T-1}, c_T, r_T, s_T$ をエピソードする．ただし，$s_0, \cdots, s_{T-1} \in S - S^+$，$s_T \in S^+$ である．このエピソードに対し総報酬から総コストを引いた

$$R' = r_1 - c_1 + r_2 - c_2 + \cdots + r_T - c_T \quad (3)$$

$$= \sum_{k=1}^{T}(r_k - c_k) \quad (4)$$

を利益と定義する．

3. 従来アルゴリズム

ここでは，まず割引収益を最大化することを目的とする Q-learning[Watkins 92] を紹介する．この手法は割引収益を最大化する問題において各状態行動価値が最適解への収束性を保証している．

次に，割引なしの場合，Q-learning よりも性能がよいとされる R-learning[Schwartz 93] を紹介する．本研究では割引なしの場合を考えるため，この手法も評価対象とする．

Profit Sharing[Holland 86, Grefenstette88] を紹介する．この手法は終端状態でのみ報酬が発生する問題においてエピソード終了時に今までの状態行動対の系列に対して一括更新を行うので学習伝搬が早いという特徴を持っている．

3.1 Q-learning アルゴリズム

Q-learning[Watkins 92] は，報酬に至るエピソードの各ステップごとに，以下の式 (5) を用いて，状態と行動の各組に対する Q 値を更新することを繰り返して政策を形成する．

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \quad (5)$$

ここで，γ $(0 \leq \gamma \leq 1)$ は割引率，α $(0 < \alpha \leq 1)$ は学習率である．この手法は，状態行動の価値が最適解へと収束することを保証している．

3.2 R-learning アルゴリズム

R-learning[Schwartz 93] は経験が，有限な収益を持つようなエピソードに分割されることがなく，割引が行われない場合の強化学習問題を扱うための手法である．この場合の目標は単位時間ステップあたりの報酬の最大化である．政策 π に従うときの単位ステップあたりの平均期待報酬に比例する値として，次式のような政策の価値関数を定義する．

$$\rho^\pi = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} E_\pi\{r_t\} \quad (6)$$

この価値関数を使って各状態行動での平均報酬を等しくする．

3.3 Profit Sharing アルゴリズム

Profit Sharing[Holland 86, Grefenstette 88] は，経験を強化する代表的な強化学習アルゴリズムであり，それぞれの状態における行動の優先度を学習する．状態 s における行動 a の優先度 $P(s, a)$ に応じて行動を選択する．

Profit Sharing は，エピソードに含まれる各状態行動対 s_t, a_t の対からなる系列を記憶しておき，報酬が得られた時点で一括して系列上の優先度 $P(s, a)$ を次式に従って強化する．

$$P(s_t, a_t) \leftarrow P(s_t, a_t) + f(t, r_T, T) \quad (7)$$

ここで，f は**信用割当関数** (credit assignment function) と呼ばれる関数である．これは，等比減少関数がよく使われているため，本研究ではこの等比減少関数を改良してモデルに適用させる．ある一つの状態行動対 s, a について，エピソードに出現した信用割当関数の値を合計したものを**強化値**と呼び，$P(s, a)$ と表す．

表 1: Q-learning 先送り更新

$Q(s, a)$ を任意に初期化
各エピソードに対して繰り返し：
  s を初期化
  エピソードの各ステップに対して繰り返し：
    Q から導かれる政策を使って，状態 s から
    行動 a を選択する
    行動 a を取り，c, r, s' を観測する
    もし終端状態でなければ，
      $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[\gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)]$
    もし終端状態ならば，
      $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[\sum_{k=1}^{T}(r_k - c_k) + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)]$
    各ステップに対する繰り返しを終了

4. コスト付き MDP への適用

ここでは，先に提案したコスト付き MDP へ，従来アルゴリズムを適用する方法を考えアルゴリズムを拡張する．適用方法は報酬とコストを終端状態まで先送りし，そのエピソードで獲得した利益によって更新をおこなう方法，コストを負の報酬として更新する方法，Profit Sharing の信用割当関数を変更する方法を考える．

4.1 利益による更新

利益最大化を目標とするため，更新もそれにともない報酬とコストを終端状態まで先送りしてそのエピソードで獲得する利益によって更新をおこなう（先送り更新）方法が考えられる．これは、コストが発生する環境でのエピソード $s_0, a_0, c_1, r_1, s_1, a_1, c_2, r_2, \cdots, r_T, s_T$ を，コストなしのエピソード $s_0, a_0, r_1(=0), s_1, a_1, r_2(=0), \cdots, r_T(=\sum_{k=1}^{T}(r_k - c_k)), s_T$ と見なすことに相当する．

Q-learning にこの更新方法を適用した場合を表 1 に示す．

Profit Sharing では終端状態でのみ報酬発生とした場合，コストを先送りして利益によって更新する．この場合，利益を時間によって割引していくように変えるため信用割当関数を利益等比減少関数と呼ぶ．

この更新方法を表 2 に示す．

4.2 コストを負の報酬として更新

利益を最大にするためにはコストを最小におさえる必要があるためコストを負の報酬として考える．これは報酬とコスト

表 2: Profit Sharing 利益等比減少関数更新

$P(s,a) = C$ と任意に初期化 (C は任意の小さな正の定数)
各エピソードに対して繰り返し：
　s を初期化
　エピソードの各ステップに対して繰り返し：
　　P に比例した確率分布に従って，状態 s から
　　行動 a を選択する
　　行動 a を取り，c, r, s' を観測する
　　$s \leftarrow s'$
　s が終端状態ならば繰り返しを終了
　エピソードに含まれるすべての状態行動対に対して：
　　$P(s_t, a_t) \leftarrow P(s_t, a_t) + \gamma^t \sum_{k=1}^{T} (r_k - c_k)$

表 3: Q-learning 逐一更新

$Q(s,a)$ を任意に初期化
各エピソードに対して繰り返し：
　s を初期化
　エピソードの各ステップに対して繰り返し：
　　Q から導かれる政策を使って，状態 s から
　　行動 a を選択する
　　行動 a を取り，c, r, s' を観測する
　　$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} - c_{t+1} +$
　　　　$\gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)]$
　　$s \leftarrow s'$　　s が終端状態ならば繰り返しを終了

が発生したらそのたびに更新をおこなっていく（逐一更新）．Profit Sharing では負の報酬が扱えないためこの方法は採用しない．

Q-learning にこの更新方法適用した場合を表 3 に示す．

4.3 Profit Sharing 信用割当関数の変更

等比減少関数は時間による割引をおこなっているが，利益最大化に対しては時間による概念は必要としない．そのためコストによる割引を考える．

$$f(t, r_T, T) = \gamma^{\delta \sum_{k=t+1}^{T} c_k} r_T \quad (0 \leq \gamma \leq 1) \tag{8}$$

この関数を収益コスト比減少関数と呼ぶ．

この更新方法を表 4 に示す．

5. 実験と結果

5.1 実験

実験 1

本研究で提案したコスト付きマルコフ決定過程モデルに対し提案したアルゴリズムを用いて実験をおこなった．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 0.04 （コスト比 小）と 0.12（コスト比 大）掛かる場合でおこなった．行動選択は Q-learning, R-learning

表 4: Profit Sharing 収益コスト比減少関数更新

$P(s,a) = C$ と任意に初期化 (C は任意の小さな正の定数)
各エピソードに対して繰り返し：
　s を初期化
　エピソードの各ステップに対して繰り返し：
　　P に比例した確率分布に従って，状態 s から
　　行動 a を選択する
　　行動 a を取り，c, r, s' を観測する
　　$s \leftarrow s'$
　s が終端状態ならば繰り返しを終了
　エピソードに含まれるすべての状態行動対に対して：
　　$P(s_t, a_t) \leftarrow P(s_t, a_t) + \gamma^{\delta \sum_{k=t+1}^{T} c_k} r_T$

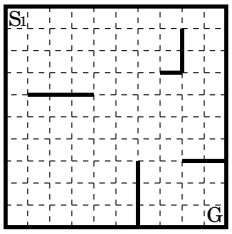

図 1．迷路問題

では Boltzmann 選択を，Profit Sharing ではルーレット選択をとった．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）として横軸に学習したエピソード数をとり縦軸にそのエピソードでの利益をとったものを図 2, 3, 4, 5 に示す．パラメータは $\gamma = 0.9, \alpha = 0.1, \beta = 0.1$，Boltzmann 選択における温度 $t = 0.2$ とする．

実験 2

次に同じ環境で Profit Sharing と Q-learning 逐一更新のアルゴリズムで γ の値をコストがよりよく使われるように設定する．Profit Sharing 利益等比減少関数，Q-learning 逐一更新では $\gamma = 1.0$，Profit Sharing 収益コスト比減少関数では $\gamma = 0.5$ で実験をおこなう．コスト比 大 で探査と知識利用による結果を図 6, 7 に示す．パラメータは $\alpha = 0.1, \beta = 0.1, t = 0.2$ とする．

5.2 実験による検討

実験 1

実験 1 の結果から検証する．まず Q-learning，R-learning の先送り更新についてだがこれはエージェントが経験した各エピソードに対して利益が異なってくる．これはエピソードが変われば総コストの量が変わるためである．このことにより環境がマルコフ決定過程を満たしていないことになる．このた

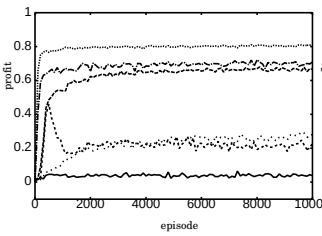

図 2．探査（コスト比 小）

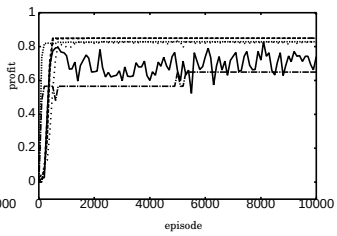

図 3．知識利用（コスト比 小）

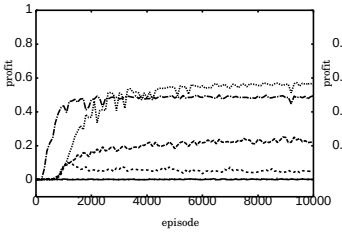

図 4. 探査（コスト比 大）

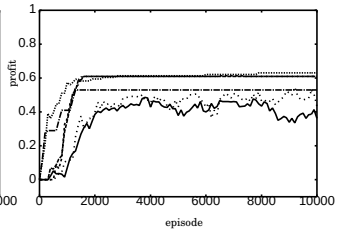

図 5. 知識利用（コスト比 大）

め状態行動対の最適な値が決まらず，結果が悪くなったと思われる．

Q-learning，R-learning の逐一更新は知識利用の結果においてよい結果となった．Q-learnig の探査においてはあまりよい結果とならなかった．

Profit Sharing はすべての場合において探査，知識利用の両方においてよい性能を示した．収益コスト比減少関数はすべてのアルゴリズムに対して一番安定した結果となっている．これはコストに対して最も大きく変化がでるためと思われる．

また総合的にみて，$\gamma = 0.9$ で統一した実験に対して探査，知識利用ともによいのは Profit Sharing に収益コスト比減少関数を用いたものである．

実験 2

Profit Sharing 利益等比減少関数の $\gamma = 1.0$ のときは性能は実験 1 と比べて，あまり変化が見られなかった．

Profit Sharing 収益コスト比減少関数の $\gamma = 0.5$ では，強化の割合が小さくなっているため探査において学習が遅くなるが，知識利用では $\gamma = 0.9$ よりよくなっている．これはコストによる強化値の変化が大きくなったためである．

逐一更新 Q-learning の $\gamma = 1.0$ のときは $\gamma < 1$ よりも利益最大化問題に対してよい性能を示し，最適の価値を求めることができるであろうことがわかった．これはコストが負の報酬として与えられることによって，割引がなくてもコストが割引の役割を果たしているからである．ただし，$\gamma = 1.0$ のとき Q-learning は最適解へ収束することが保証されていない．

6. まとめと今後の課題

6.1 まとめ

本研究では強化学習に対して今までの収益最大化ではなく，新たに利益最大化問題を提案し，これを強化学習アルゴリズムを用いて解く方法を実験によって検討した．

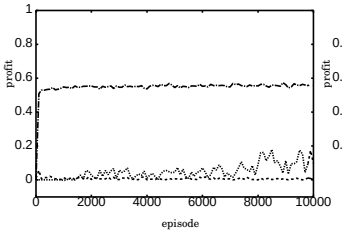

図 6. 探査（コスト比 大）

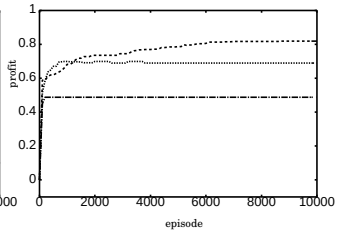

図 7. 知識利用（コスト比 大）

表 5: 利益最大化問題における各アルゴリズムの性能評価

	学習中	知識利用	学習速度
QL 先送 $\gamma = 0.9$	×	×	×
QL 逐一 $\gamma = 0.9$	△	△	△
RL 先送 $\gamma = 0.9$	×	×	×
RL 逐一 $\gamma = 0.9$	△	△	△
PS 利益等比 $\gamma = 0.9$	○	△	◎
PS 収益コスト比 $\gamma = 0.9$	◎	○	◎
QL 逐一 $\gamma = 1.0$	×	◎	△
PS 利益等比 $\gamma = 1.0$	○	△	◎
PS 収益コスト比 $\gamma = 0.5$	×	○	△

注：◎は最も優れているものを，○は優れているものを，△はあまり優れていないものを，×は悪いものを表す．

利益最大化問題に対しての各アルゴリズムの性能を評価したものを表 5 に示す．

Q-learning, R-learning の先送り更新では環境がマルコフ性を満たさず，実験でもよい結果とならなかった．

今回の利益最大化問題に対しては，学習中では Profit Sharing の収益コスト比が最も良い性能を示したのでこの方法を使うべきである．また知識利用では Q-learning の逐一更新での $\gamma = 1.0$ が最も良い性能を示したのでこの方法を使うのがよいことがわかった．学習に時間をかけれない場合には，収束の早い Profit Sharing を使うべきである．

6.2 今後の課題

本研究では利益最大化問題に対して今までの手法を変更してアルゴリズムの性能を評価してきたが最適利益への収束は保証されていない．既存のアルゴリズムは割引収益最大化をするための手法であるため，新たにアルゴリズムを考える必要がある．

また，今後はこの利益最大化問題に対して状態遷移の不確定性や不完全知覚問題などのエージェント自身の問題などの性能についても調べる必要がある．

参考文献

[RL 98] Richard S.Sutton , Andrew G.Barto: Reinforcement Learning, MIT press(1998)

[Watkins 92] Watkins, C.J.C.H., and Dayan, P.: Technical note: Q-learning, Machine Learning, Vol.8, pp.55-68(1992)

[Schwartz 93] Anton Schwartz: A reinforcement learning method for maximizing undiscounted rewards, In Proceedings of the Tenth International Conference on Machine Learning, pp298–305. Morgan Kaufmann(1993)

[Holland 86] John H Holland: Escapaing brittleness:The possibilities of general-purpose learning algorithms applied to parallel rule-based system, Machine Learning:An Artificial Intelligence Approach, Vol.2(1986)

[Grefenstette 88] John J Grefenstette:Credit assignment in rule discover systems based on genetic algorithms, Machine Learning, Vol.3, pp225-245(1988)