RAE-PIA: 複数報酬環境下における最適政策の効率的強化学習

山口　智浩*・天正　新二郎**・石村　健二***

RAE-PIA: An efficient reinforcement learning method for an optimal policy under a multi-rewards' environment

Tomohiro Yamaguchi, Shinjiro Tensho and Ishimura Kenji

This paper presents a new model based reinforcement learning method that directly updates Reward Acquisition Efficiency by the step (RAE, in short) in each rule. Most reinforcement learning methods optimize the discounted total reward received by an agent. However, discounting encourages the learner to sacrifice long-term benefits for short-term gains. Besides, using small discount factor leads to sub optimal policy, on the other hand, larger discount factor makes the learning time so large. To solve them, in our method, while identifying the environment by MDP model, it estimates the sum of received rewards and an expected distance of the reward acquisition loop with the model to integrate them into the RAE.

1　はじめに

人工知能の主要な研究分野の１つに強化学習[1,2,3]がある。**強化学習**（Reinforcement Learning、以下**RL**と略記）とは、環境からの感覚入力に対するエージェントの行為（感覚・行為の組み合わせを**行動**と呼ぶ）に対して環境から受け取る**報酬**（**Reward**）のみを手がかりとして、将来得ることができる単位時間当りの報酬和を最大にする、各状態ごとの行動の集合（**政策**と呼ぶ）を獲得する方法である。

強化学習での本来の学習基準は、学習エージェントが受け取る１ステップ当たりの**平均報酬**の最大化である。これを直接求める手法は、平均報酬強化学習法[4,5]と呼ばれる。しかしながらこの平均報酬に基づく最適化手法では、平均報酬が受け取る報酬の無限ステップ和を用いて定義されるため、平均報酬の見積もりが理論的[4]にも手続き的[5]にも困難である、という問題点[4,5]があった。

この問題点は、意志決定理論での“**割引率**”を用いて将来に渡る無限の報酬和を現在から遠ざかる報酬ほど割り引き、数学的に収束する報酬の有限和を表わす**効用値**（utility）として見積もることで回避できる。そのため、最適政策を求める大半の強化学習研究は、**割引期待報酬和**の最大化[1,6,7]を学習基準としている。しかしながら割引型最適化手法では、１）学習結果（収束政策）と学習速度とが用いる割引率の値に依存し、かつ最適政策の質と学習速度とが両立せず、しかも２）割引率の合理的な値決定法がない、という２つの基本的な問題点があった。

各状態の効用値の収束計算法の代表が、**政策反復アルゴリズム**（Policy Iteration Algorithm:PIA）[1,6]である。しかしながら**PIA**には、状態数が多くなると収束に要する反復回数が組合せ的に増大する、という基本的問題がある。これは、効用値更新の際に、各状態の効用値間の依存関係を考慮することなく一律の順序で全状態の効用値更新（網羅的再計算）の反復計算を行うからである。

この欠点を解決する手法として**優先掃き出し法**（Prioritized Sweeping）[7]があり、優先度を用いて効用値間の依存関係を推定し、PIAの効率化を図っているが、学習速度や更新回数には、改善の余地がある。

そこで本論文では、平均報酬強化学習法に代表される**非割引型最適化**[4,5]の新手法として、報酬からの期待距

＊情報工学科
＊＊電子情報専攻科
＊＊＊現北陸先端科学技術大学院大学

注１）人工知能学会全国大会（第15回）
2001年6月24日、口頭発表の論文を加筆修正

離に着目し、報酬獲得効率（Reward Acguisition Efficiency: RAE）を最大化する手法：**RAE-PIA**を提案[8,11]し、比較学習実験結果を報告する。決定的MDP環境において、本手法を複数報酬環境に拡張したアルゴリズムを用いて、割引型PIA[1]、Prioritized Sweeping[6]とで最適政策の学習コストを比較実験した結果、従来手法と比べて大幅な学習コストの減少を実証した。

2　MDPモデルに基づく強化学習法

本論文で対象とする**モデルに基づく強化学習法**（model based RL）[6]は、最適政策の学習を、1）観測による環境モデル同定と、2）モデル上での最適政策の探索、との2段階に分割して行う点が特徴である。

2.1 政策反復アルゴリズム（Policy Iteration Algorithm：PIA）[1,6]

PIAとは、環境モデル上において各政策での割引期待報酬和を比較しながら政策を更新し、最適政策に収束するまで各状態の割引期待報酬和を反復計算し、状態の効用値を最大化する政策を求めるアルゴリズムである。しかしながら、PIAは各状態の効用値の依存関係を考慮せずに効用値の変動のない状態を含めた全ての状態について一律の順序で網羅的に再計算（sweep）するため、不必要な効用値更新が多く含まれ、状態数が大きい程非効率で計算コストが大きくなる。さらに、割引型PIAでは学習結果の質と学習速度が**表1**に示すように割引率に関してトレードオフの関係にあり、両者を同時に改善するのは困難という問題点があった。

表1　割引型最適化における割引率と収束政策との関係

割引率	収束する政策の質	学習速度
r ≒ 1	最適政策（に近い）	遅い
r ≒ 0	準最適政策	速い

2.2 優先掃き出し法(Prioritized Sweeping)

PIAの反復計算の非効率性を改善するための手法であるPrioritized Sweeping[7]は、各状態の効用値の依存関係を考慮し、計算順序を効率化して効用値を繰り返し更新する。効率改善の基本的な考え方は、“効用値が影響を受けうる状態に正の優先度を割り当て、優先度の大きな状態から優先的に効用値を更新し、優先度が0の状態の更新を行なわない”、ということである。実際に学習コストが改善されるかは、効用値更新回数の減少分と優先度算出コストとの差によって決まる。

Prioritized Sweepingのアルゴリズムは動作を大きく分けて、効用値更新部と優先度更新部の二つに分けられる。効用値更新部では、優先度更新部で求められた状態について各状態の割引期待報酬和を効用値として更新を行う。優先度更新部では、各状態の計算順序の優先順位を決定する優先度を求める。その後、優先度の一番大きい状態について効用値、及び優先度を計算し、これを繰り返す。式(1)に優先度Pr(i')を求める式を示す。

$$\Pr(i') = M^{a'}_{i'i} \cdot \Delta U(i) \qquad (1)$$

Pr(i'):状態i'の優先度，$M^{a'}_{i'i}$:現状態iへ状態i'から行為a'を行なったときの遷移確率，⊿U(i):更新前後の効用値の差分

2.3 報酬獲得効率（Reward Acquisition Efficiency：RAE）

報酬獲得効率：RAE[8]とは、状態iで行為aを行う任意の行動ルールRule(i, a)を実行したときに、単一報酬環境下で将来得られる1ステップ当りの期待報酬を表す。式(2)にその定義を示す。

$$\text{RAE(i,a)} = \text{報酬}／\text{報酬までの最小期待距離} \qquad (2)$$

報酬までの最小期待距離の算出は、PIAと同様に繰り返し計算によって算出できる。この繰り返し計算法を**単純RAE-PIA**と呼ぶ。割引型PIAとの違いは、割引率というパラメータがなく、収束に要する繰り返し数が割引型PIAと比べ少ない回数で済む点であるが、単純RAE-PIAでの繰り返し計算は、一般的なPIAと同様に、各状態の効用値間の依存関係を考慮することなく一律の順序で反復計算を行うので、依存関係を利用した効率化が可能である。そこで、**BP法**（Back Propagation:逆伝播法）[9]を用いて、報酬を獲得したルールをルートとして遷移を逆向きに展開して、状態をノード、ルールをアークとした有向木を作成する。**図1**に報酬獲得ループを含むMDP環境の例を、**図2**にBP法による報酬獲得ルールを根とする有向木の生成を示す。図1、2の太線部は、探索された最適政策を表す。この作成した有向木を用いて状態間の効用値の依存関係を考慮した効率的な算出を行う手法を、**優先RAE-PIA**と呼ぶ。

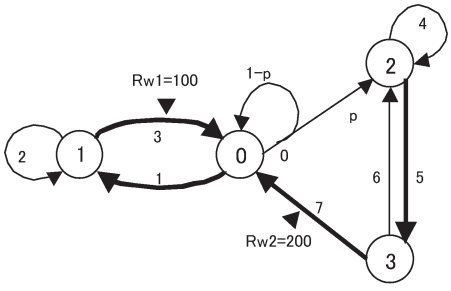


図1　報酬獲得ループを含むMDP環境の例

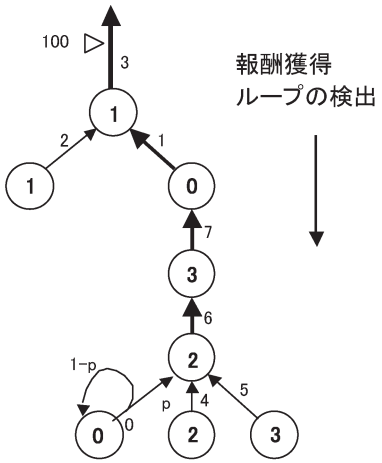


図2　BP法による報酬ルールを根とする有向木の生成

2.4　RAE-PIAの複数報酬環境への拡張

一般にMDP環境における最適政策の形状は、繰り返し報酬を獲得するループ状の状態遷移系列（**報酬獲得ループ**と呼ぶ）上における各状態の政策と、報酬獲得ループに合流する状態遷移系列上の各状態の政策との和集合で構成される。ここで報酬獲得ループのRAE値：Loop-RAEを(3)式で定義する。

Loop-RAE＝ループ上の報酬和／ループの期待距離 (3)

単一報酬環境の場合、**最適報酬獲得ループ**は、報酬ルールを含む報酬獲得ループのうち、ループの期待距離が最小となるものである。前節のBP法では、報酬ルールを起点とした横型探索により、報酬獲得ループを網羅的に探索しつつ、かつ期待距離が最小のループを効率よく検出できる。

複数報酬環境では、まず報酬を1つ以上含む報酬獲得ループを全て検出し、1）Loop-RAEが最大となる**最適報酬獲得ループ**を選び、2）最適報酬獲得ループ上の状態以外の状態について、最適報酬獲得ループに合流する（部分）最適政策を優先RAE-PIAで求める。そして，1）2）を統合することにより、複数報酬環境下での最適政策を算出する。

3　実　　験

実験1として、**図3**から**図7**に示す大きさの異なる5種類の決定的MDP環境を用いて、割引型PIA、Prioritized Sweeping [10]とRAE-PIAとの学習時間及び、効用値更新回数についての学習コストの比較学習実験を行った。**実験2**として、**図8**に示すルール数52の決定的なMDP環境で報酬数がそれぞれ1、2、4個と異なる環境について実験1と同様の項目について比較学習実験を行った。なお、実験1、2の割引型PIAとPrioritized Sweepingで用いた割引率は、0.99である。

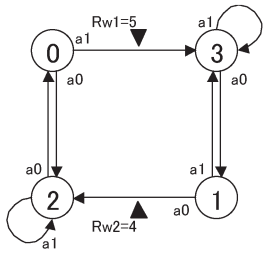


図3　実験環境1-1：4状態2行動8遷移2報酬の環境

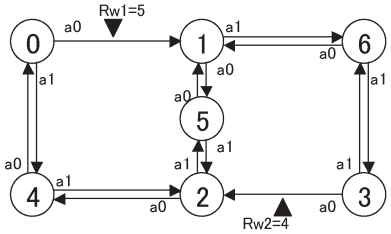


図4　実験環境1-2：7状態2行動14遷移2報酬の環境

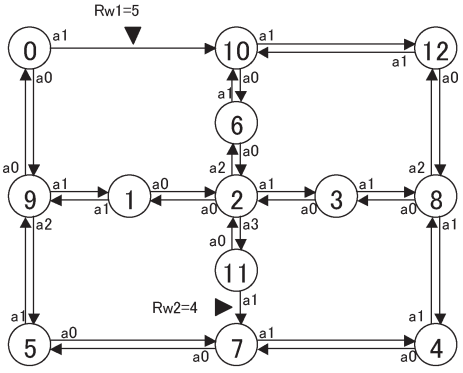


図5　実験環境1-3：13状態4行動52遷移2報酬の環境

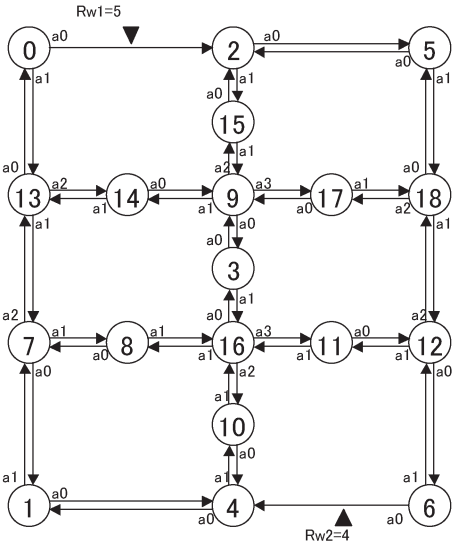


図6　実験環境1-4：19状態4行動76遷移2報酬の環境

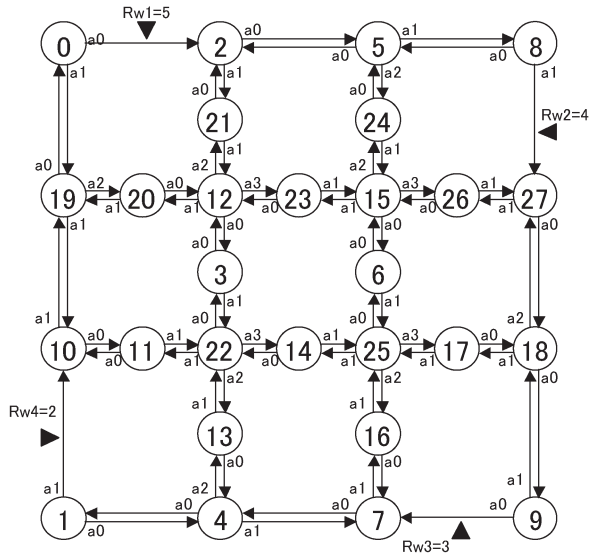


図7　実験環境1-5：28状態4行動112遷移4報酬の環境

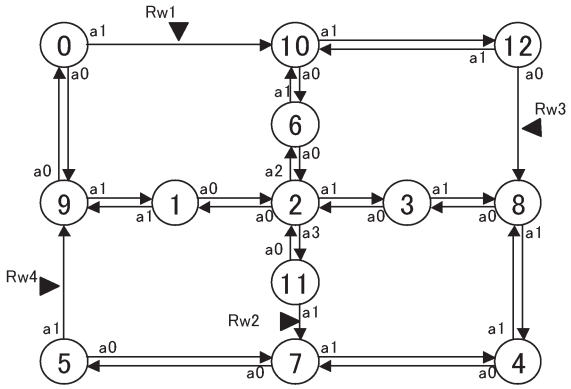
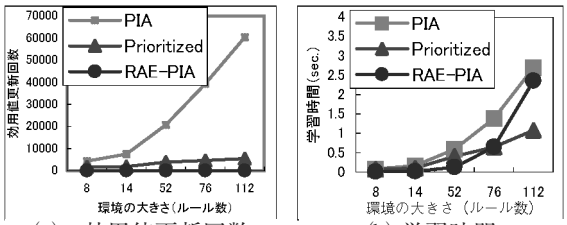


図8　実験環境2：報酬数の異なる環境での比較学習実験

4　実験結果

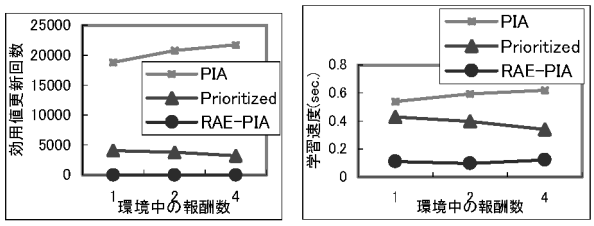
実験1，2の実験結果を、それぞれ図9、図10に示す。

実験1：ルール数の違いによる学習コスト比較では、図9(a)効用値更新回数は、PIAの更新回数がルール数に対して増加しているのに対し、Prioritized Sweeping とRAE-PIAは大きく改善されている。図9(b)学習時間のルール数に対する増加は、PIAが最も大きく、Prioritized Sweeping 、RAE-PIA、の順に小さくなっている。ルール数76以下でPrioritized Sweepingの学習時間がRAE-PIAより大きいのは、優先度の更新コストのためと考えられる。一方、ルール数112でRAE-PIAの学習時間が悪化しているのは、報酬数が2から4と2倍に増加し、RAE-PIAの計算量が報酬数に依存するためと、現状のRAE-PIAにおいて報酬ごとに有向木の展開を行なうアルゴリズムの実装レベルでの冗長さが原因である．



(a)　効用値更新回数　　(b) 学習時間

図9　実験1：ルール数の違いによる学習コスト比較



(a)　効用値更新回数　　(b) 学習時間

図10　実験2：報酬数の違いによる学習コスト比較

実験2：報酬数の違いによる学習コスト比較では、図10(a)効用値更新回数、図10(b)学習時間とも、PIAが最も大きく，Prioritized Sweeping 、RAE-PIA、の順に小さい。PIAは、環境中に含まれる報酬が増えると、学習コストが増加するのに対し、逆にPrioritized Sweepingは、報酬が多いほど減少する。これは、報酬が多い程、優先度の計算が効果的に効用値更新場所と順序を制御するからだと考えられる。これに対しRAE-PIAは、前述の理由で報酬が増えるごとに実装に依存した学習時間が若干増すが、他に比べると論理的な学習コストを意味する効用値更新回数が格段に少なく安定している。

5　まとめ

本論文では、MDPモデルに基づく非割引型強化学習手法としてRAE-PIAを提案し、本手法が実験環境の大きさや報酬の数に関係なく、PIAやPrioritized Sweepingと比べ学習時間が短く済み、効用値更新回数も格段に少ない回数で学習できることを実験的に示した。本実験で用いたRAE-PIAのアルゴリズムは、4章の実験結果で述べた通り改善の余地が残されている。具体的には、各報酬ごとに各状態から報酬にたどり着くまでの経路を木に展開しているため、全ルールが報酬の数だけ重複して探索されているのが原因である。今後の課題は、本提案アルゴリズムの確率的環境への適用及び、RAE-PIAの効率改善手法の実装である。

参考文献

[1] Russell, S. J., and Norvig, P., AI -a Modern Approach-, Prentice-Hall International, Inc., pp.598-624, 1995

[2] Sutton, R. S., and Barto, A. G., Reinforcement Learning: An Introduction, MIT Press, Cambridge, MA, 1998

[3] Kaelbling, L. P., Littman, M. L., and Moore,A. W., Reinforcement Learning: A Survey, Journal of Artificial Intelligence Research, Vol.4, pp.237-277, 1996

[4] Mahadevan, S., Average Reward Reinforcement Learning: Foundations, Algorithms, and Empirical Results, Machine Learning, 22, pp.159-195, 1996

[5] Tadepalli, P. and Ok, D., Model-based Average Reward Reinforcement Learning, Artificial Intelligence 100 , pp.177-224, 1998

[6] Barto, A. G., Bradtke, S. J. and Singh, S. P., Learning to act using real-time dynamic programming, Artificial Intelligence 73, pp.81-138, 1995

[7] Moore, A. W., Atheson, C. G., Prioritized Sweeping: Reinforcement Learning With Less Data and Less Time, J. of Machine Learning 13, pp.103-130, 1993

[8] 山口 智浩，石村 健二，RAE-PIA：報酬獲得効率を最大化する政策の強化学習，人工知能学会全国大会論文集，pp.329-332，2000

[9] 谷元 昭文，山口 智浩，逆伝播法による報酬獲得効率の高速かつ柔軟な強化学習手法の提案，電気学会関西支部高専卒業研究発表会論文集，pp.25-26，2000.3

[10] 天正 新二郎，山口 智浩，優先掃き出し法による最適政策の効率的な強化学習，電子情報通信学会 学生研究発表会論文集，p.98，2001.3

[11] 山口 智浩，石村 健二，RAE-PIA：報酬獲得効率を最大化する政策の強化学習，奈良高専研究紀要，vol.36, pp.101-106，2000