論　文

動的環境下における Profit Sharing

加藤　新吾[†]　　松尾　啓志[†]

A Theory of Profit Sharing in Dynamic Environment

Shingo KATO[†] and Hiroshi MATSUO[†]

あらまし　マルチエージェント系や実環境において自律的に環境に適応していく学習方法として強化学習が注目されている. このような問題を対象としたとき学習器の環境に変化が生じると考えることは自然である. 本論文では, はじめに環境の変化後に Profit Sharing が無効ルールを抑制するのに必要なエピソードについて解析的に考察した. その後, 従来法よりも高速に無効ルールを抑制することが可能な忘却型 Profit Sharing を提案し, 実験によりその有効性を確かめた.

キーワード　強化学習, profit sharing, 動的環境

1. はじめに

強化学習 (reinforcement learning) [1] とは, 報酬 (reward) という特別な入力を手掛かりとして環境に適応する教師無し機械学習の一種である. 強化学習の目的は, できるだけ多くの報酬をできるだけ早く獲得することである. 最終的に多くの報酬を得るために最適性を重視した接近を環境同定型と呼び, 学習途中でも報酬を得る効率性を重視した接近を, 経験強化型と呼んでいる [2]. 宮崎ら [3] は, 報酬獲得と環境同定のトレードオフを考慮し, 報酬獲得器と環境同定器を切り替える強化学習システムを提案した. また片山ら [4] は, 最適化ではなく満足化, つまり一定の水準を達成することを最終目的とする学習システムを提案した.

多くの強化学習とその理論的な解析は, 環境をマルコフ決定過程 (Markov Decision Processes : MDPs) としてモデルしており, 状態の観測は完全であることを仮定していた. 例えば宮崎ら [5] によって, 離散マルコフ決定過程での強化学習法のいくつかが紹介されている. 最近では部分観測マルコフ決定過程 (Partially Observable MDPs : POMDPs) を問題領域とした研究も進められており, 木村ら [6] はいくつかの典型的な強化学習に対する概説を行った. 他にも, 決定的政策により合理性が保証される POMDPs を対象とし, 重みを使わずに合理的政策を形成する手法である合理的政策形成アルゴリズムが提案された [7].

強化学習を, ロボットへ応用する研究 [8] も行われている. 例えば, ロボットの歩行制御への適用 [9] や, 人間からの助言を政策に反映できるシステムを提案し, それを実際のロボットに組み込んだ場合の評価が行われた [10].

また最近では, 複雑かつ動的な問題を多くの自律的なエージェントが何らかの協調動作をすることで解決しようと研究がすすめられており, マルチエージェント系における協調行動の実現は, 工学的及び認知科学的観点から興味ある話題である. しかし, これらの多くは人間が設計した行動群によって制御されており, 対象やタスク, 個体間の相互作用が複雑になるつれてその設計が困難になる. そのため, エージェント自身の学習や適応能力が求められている. このような背景から強化学習法が注目されている. 例えば, 他のエージェントの行動を模倣することにより学習速度を向上させる研究が行われている [11]. また荒井ら [12] は, 環境同定型の代表的手法である Q-learning と経験強化型の代表的手法の Profit Sharing を取り上げ, マルチエージェント系に対する代表的なベンチマーク問題である追跡問題を通じて, マルチエージェント強化学習として Profit Sharing の優位性を主張している. 宮崎ら [13] はマルチエージェント強化学習として Profit Sharing を用いた場合で, 単位行動当りの期待獲得報酬を正にするための各エージェントへの報酬配分に関する必用

[†] 名古屋工業大学工学部, 愛知県
Nagoya Institute of Technology, Gokiso-cho, Showa-ku, NAGOYA, 466–8555, JAPAN

十分条件を導出した. エレベータ問題へマルチエージェントの強化学習を適応することにより, 従来のアルゴリズムよりも優れた性能が出る事が Robert [14] らによって確認された.

強化学習をより多くの問題に適用するためには, 数学的な解析が重要な役割を担うと考えられる. 強化学習の解析的な研究の多くは環境に変化が生じないことを前提としているが, 実環境やマルチエージェント系では外乱や故障などにより環境が変化すると考えるのが自然である. 例えば, ロボット制御問題の場合, 現実世界ではモータの磨耗や故障が生じないと考えるのは不自然である. また, マルチエージェント環境では, 各エージェントが独立して学習を行うため, 学習毎に環境の遷移確率が変化すると考えることができる.

確率的傾斜法 [15] を用いて, 内部変数の変化を調べることにより環境変化を識別する研究 [16] が行われた. しかし, 次の環境変化がいつ起こるかは一般に予測することはできないため, 変化後の環境に学習器をどのように適応させるかという問題も残っている. 動的に変化するネットワークに対して強化学習を適応させる研究 [17], [18] も行われている.

本論文では,Profit Sharing の環境変化が生じた場合における無効ルール抑制に対する理論的な考察を行う. その後に, 高速に無効ルールを抑制できる忘却型 Profit Sharing を提案する. 以下 2 章では, 準備として Profit Sharing 学習法の合理性定理と, 環境に変化が生じた場合における従来法の特性について論じる. 3 章では提案手法について論じる.4 章では, 従来法と提案手法の比較実験について, 実験の方法, 実験の結果及び考察を述べる.5 章は結論で, 本研究の成果をとりまとめ, 今後の研究課題について論じる.

2. Profit Sharing

2.1 準　　備

Profit Sharing は, 報酬に至るエピソードにおける感覚入力 x と行動 a の対からなるルール系列を記憶しておき, 報酬が得られた時点で系列上のルールを次式に従って強化する.

$$w(x_i, a_i) \leftarrow w(x_i, a_i) + f(r, i) \qquad (1)$$

ここで,$w(x_i, a_i)$ はエピソード系列上の i 番目のルールの重み,r は報酬値, f は強化関数である.

感覚入力 x で行動 a を選択する"if x then a" というルールを $\overrightarrow{xa}$ と書く. あるエピソードで, 同一の感覚入力に対して異なるルールが選択されているとき, その間のルール系列を迂回系列 (detour) と呼ぶ. 例えば, 図 1 の環境で, エピソード ($\overrightarrow{yb} \cdot \overrightarrow{xb} \cdot \overrightarrow{zb} \cdot \overrightarrow{yb} \cdot \overrightarrow{xa}$) には, 迂回系列 ($\overrightarrow{xb} \cdot \overrightarrow{zb} \cdot \overrightarrow{yb}$) が存在する (図 2). 迂回系列上のルールは, 報酬の獲得に貢献しない可能性がある. 現在までのエピソードで, 常に迂回系列上にあるルールを無効ルール (ineffective rule) と呼び, それ以外を有効ルール (effective rule) と呼ぶ. 無効ルールと有効ルールとが競合するならば, 明らかに無効ルールを強化するべきではない.

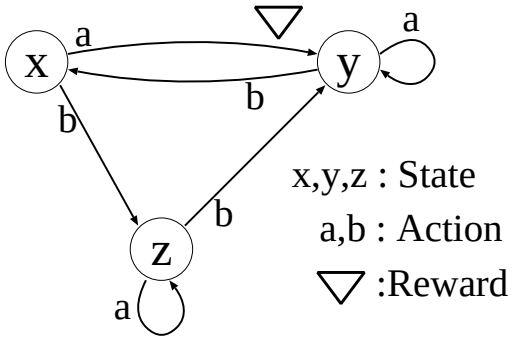


図 1　環 境 の 例
Fig. 1　Example of environment

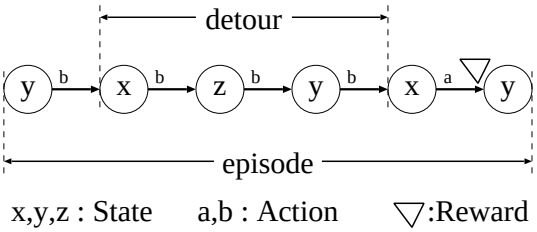


図 2　エピソードと迂回系列の例
Fig. 2　Example of episode and detour

無効ルールの重みが最大でないとき, 無効ルールは抑制されていると呼ぶ. 次式を満たす強化関数であれば, 有効ルールは無効ルールを抑制することができることが合理性定理により証明されている [19].

$$L \sum_{j=i}^{W} f(r, j) < f(r, i-1), \quad (\forall i = 1, 2, \cdots, W.) \qquad (2)$$

ここで,W はエピソードの最大長,L は同一感覚入力下に存在する有効ルールの最大個数である. 定理を満た

す最も簡単な強化関数として, 次に示す等比減少関数が考えられる.

$$f(r,n) = \frac{1}{S} f(r,n-1), \quad n = 1,2,\ldots,W-1. \tag{3}$$

ここで,S(但し $S \geq L+1$) を報酬割引率と呼ぶ.

合理性定理は最適性を保証していないが,MDPs の仮定を必要としないのでマルチエージェント系のような非 MDP 環境に対しても適用できる点に特徴がある.

今後, 理論的な評価では MDPs を仮定しないものとし, 強化関数は実験を含め合理性定理を満たすものとする. また, 報酬が最大となる無効ルールと有効ルールが得る報酬の比率を P と表す.

$$P \times f(r,i-1) = L \sum_{j=i}^{W} f(r,j), \quad (P < 1) \tag{4}$$

例えば式 (3) で $W = \infty$ の時,P は次式で表される.

$$P = \frac{L}{S-1} \tag{5}$$

2.2 環境変化後に無効ルールを抑制するのに必要なエピソード数

本節では, 環境変化後に無効ルールの抑制を保証するエピソード数を導出する. 無効ルールの抑制とは, 無効ルールの重みがそれと競合する有効ルールを差し置いて最大の重みにならないことである. また, 無効ルールを抑制するのに必要なエピソード数が大きいことを抑制が困難であると表すこととする. つまり最も困難な条件とは, 無効ルールを抑制するのに必要なエピソード数が最大となる条件である.

まず, 変化後の環境下で無効ルールを抑制するのが最も難しい条件を示す. 次に, 最も困難な条件下で無効ルールを抑制する為に必要なエピソード数を求める. 最後にそれを任意の無効ルールに拡張する.

[補題 1] 最も困難な条件

現在の環境下での唯一の回帰的無効ルールが, 変化前の環境において同一感覚入力下で得られる報酬を独占していた場合に最も困難になる. □

証明は**付録 1** に示す. 図 3 に唯一の回帰的無効ルールの例を示す. ここで, 回帰的とは行動をとった結果, 感覚入力の変化が生じないルールのことをいう.

[補題 2] 最も困難な条件下で無効ルールを抑制するのに必要なエピソード数

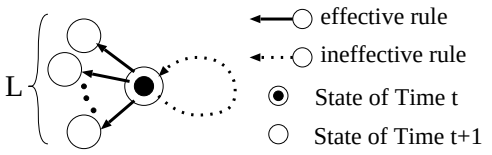


図 3 唯一の回帰的無効ルール

Fig. 3 The state with single recursive and ineffective rule.

最も困難な条件下で無効ルールが抑制される条件は,

$$\frac{L \cdot Rw'}{(1-P)Rw} \cdot G < X \tag{6}$$

である. □

証明は**付録 2** に示す. ここで,X は環境変化後のエピソード数,G は変化前の環境のエピソード数,Rw は変化後の環境で有効ルールに与えられる報酬, Rw' は変化前の環境で有効ルールに与えられる報酬である.

補題 1 と 2 から, 推移律により直ちに次の定理が得られる.

[定理 1] 無効ルール抑制に必要なエピソード数

環境変化後の任意の無効ルールが抑制される条件は

$$\frac{L \cdot Rw'}{(1-P)Rw} \cdot G < X \tag{7}$$

である. □

ここで,X は環境変化後のエピソード数,G は変化前の環境のエピソード数,Rw は変化後の環境で有効ルールに与えられる報酬, Rw' は変化前の環境で有効ルールに与えられる報酬である.

2.3 定理の意味

式 (7) に示すように, 新たな環境で無効ルールを確実に抑制するためには, 変化前の環境でのエピソード数 G の数倍のエピソードが必要である. たとえば $L = 3, P = 0.75, Rw = Rw'$ のとき, 確実に無効ルールを抑制するのに必要なエピソードは環境変化前の 12 倍のエピソードが必要である.

また, 参考のために有効ルールと無効ルールが入れ替わる場合での最も容易な条件についても考える. 最も困難な条件とは逆に, 変化前の環境下で唯一の回帰的無効ルールが, 変化後に得る報酬を独占する場合が最も容易に無効ルールを抑制できる. つまり, 抑制に必要なエピソード数 X の条件は次式で表される.

$$\frac{(1-P)Rw'}{L \cdot Rw} \cdot G < X \tag{8}$$

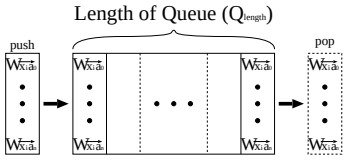
環境が変化しても有効ルールと無効ルールが入れ替わらない場合もあるが, もし入れ替わった場合は, 最適な行動を取ったとしても, 抑制に必要なエピソード数 X は環境変化前のエピソード数 G に比例して増大することが式 (8) から示される.

3. 忘却型 Profit Sharing

従来手法が環境変化に対して高速に適応できない原因として, 過去の情報を保持し続けることが考えられる. 新たな環境に適応するためには, 環境が変化すると同時に新しい環境で不利になるルール重みを忘れてしまうことが望ましい. しかし, 環境変化の識別及び新たな環境でどのルール重みが不利になるかを知ることは非常に困難である. そこで, 環境の変化を問わず常に全てのルール重みを徐々に減少していく忘却型 Profit Sharing を提案する.

3.1 提案手法 1:Profit Sharing with Queue (PSwQ)

過去の情報を徐々に消去するために, 学習器に有限長のキューを持たせ各エピソードで得た報酬をキューに入れていく手法を提案する (図 4).



$W_{\overline{xa}}$:Weight of Rule (if x then a)

図 4 ルールテーブルのキュー
Fig. 4 Schematic diagram of the rule table queue

図 4 において, キューに挿入される要素 $W_{\overline{x_i a_j}}$ は直前のエピソードでルール $\overline{x_i a_j}$ が得た報酬の和を表す. 各エピソード毎にキューは更新され, 要素の数がキューの長さを超えた場合は, 最も古い要素が消去される. また, 行動を選択する際には, キューに保持されているそのルールの重みの総和を利用する $(w(x_i, a_j) = \sum_{Q_{length}} W_{\overline{x_i a_j}})$.

[定理 2] PSwQ を用いたときの無効ルール抑制に必要なエピソード数

PSwQ を用いたとき, 環境変化後の任意の無効ルールが抑制される条件は

$$\frac{Q_{length} \cdot L \cdot Rw'}{(1-P)Rw + L \cdot Rw'} < X \qquad (9)$$

である. □

証明は**付録 3** に示す. ここで,Q_{length} はキューの長さである.

従来手法との比較:式 (7) と式 (9) とを比較すると, 式 (9) には変化前の環境のエピソード数 G が含まれていないことが分かる. つまり定理は, 無効ルールを抑制するのに従来手法では変化前のエピソード数に比例して増加するのに対して,PSwQ では定数エピソードで抑制が可能であることを示している.

3.2 提案手法 2:Profit Sharing with Virtual Queue (PSwVQ)

学習器にキューを持たせる手法は確実ではあるが, キューの長さに応じて必要なメモリが大きくなるのでキューの長さが大きくなると現実的ではない. そこで, 過去のエピソードで得た報酬の価値を低くする手法を提案する.

具体的には次式に従って強化していく.

$$w(x,a) \leftarrow w(x,a) \times \tau + \sum_{k=0}^{W} g(x,a,k) \qquad (10)$$

$$g(x,a,k) = \begin{cases} f(r,k) & (\text{if} x = x_k \text{and} a = a_k) \\ 0 & (\text{else}) \end{cases}$$

ここで,$\tau (0 < \tau < 1)$ を忘却率と呼ぶ.

[定理 3] PSwVQ を用いたときの無効ルール抑制に必要なエピソード数

PSwVQ を用いたとき, 環境変化後の任意の無効ルールが抑制される条件は

$$\tau^X < \frac{(1-P)Rw}{(1-P)Rw + L \cdot Rw' \cdot (1-\tau^G)} \qquad (11)$$

である. □

証明は**付録 4** に示す.

従来手法との比較: 式 (11) より, 変化前のエピソード数 G が増加することにより抑制に必要なエピソード数が増加することが分かる. ここで,$G = \infty$ の場合を想定すると, 無効ルールが抑制される条件は次式により表される.

$$\tau^X < \frac{(1-P)Rw}{(1-P)Rw + L \cdot Rw'} \qquad (12)$$

従って, 従来手法では式 (7) より,G に従い抑制に必要なエピソード数は無限に増加していくのに対して,PSwVQ では最悪でも定数エピソードで抑制が可能である.

4. 実　　験

本章では, 従来手法と提案手法を様々な実験環境で実装し, 無効ルール抑制に必要なエピソードなどを調べ, 結果について考察を加える.

また, 各実験においてのパラメタの設定基準において, 各ルールの重みや報酬の値に関しては浮動小数点演算の演算誤差による影響を削減するために十分大きい整数値を用い, 忘却率やキューの長さに関しては予備実験から良好な結果が得られたものを採用した.

4.1 数 値 実 験

4.1.1 実 験 方 法

図 5 に示される無効ルールが入れ替わる単純な環境を実験環境とした. 各ルールの初期値は 1000, 報酬値は 10000, 強化関数は公比 5 の等比減少関数とし, ルールの選択は重みに比例した確率に基づくルーレット選択を採用した. キューの長さを 1000, 忘却率を 0.9989 とする. 1000 エピソード経過後に環境の変化を起し, 乱数の種を変えて各々50 回ずつ計測した.

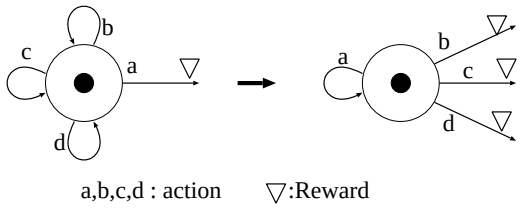


図 5　数値実験の環境

Fig. 5　Environment for numerical simulation

4.1.2 実験結果および考察

計測結果の平均値を用いた有効ルールと無効ルールの重みの比率 (抑制率) と環境変化後のエピソード数の関係を図 6 に示し, 無効ルールを抑制するためのエピソード数についての理論値と計測結果を表 1,2 に示す.

図 6 において, 抑制率が 1 を超えているとき, 無効ルールは抑制されている. また, 矢印は各手法の無効ルールを抑制するのに必要なエピソード数の理論的な範囲を示す. 定理は任意の無効ルールを抑制する (つまり抑制率が 1 を超える) のに必要なエピソードについてのみ保証しているが, 抑制後も従来手法と比較して提案手法の方が抑制率の増加が早いことが分かる.

表 1 と表 2 から実験結果が理論値の範囲内に収まっていることが分かる. また, 最小値 · 最大値に関して理論値と実験の値に差が発生した. この理由として, 理論値は最良 (最悪) な政策を選択することを想定して算出を行ったが, 一般的には政策は確率的に選択されるため, 最良 (最悪) の政策をとる可能性は低いためであると考える.

表 2 より, 提案手法 1 では従来法の 1.55 – 5.06 倍, 提案手法 2 では従来法の 1.13 – 5.27 倍の早さで無効ルールを抑制可能なことが分かる.

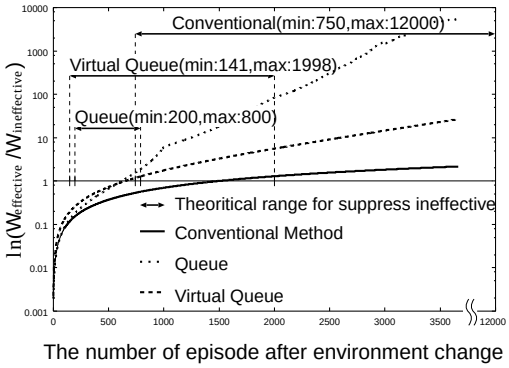


図 6　数値実験の結果; 環境変化後のエピソード数と抑制率 (有効ルールの重み／無効ルールの重み) の関係

Fig. 6　Result of numerical simulation; relation between suppression rate and the number of episode after environment change.

表 1　実験 1 での無効ルールを抑制するのに必要なエピソード数の理論値

Table 1　The theoretical result of episodes in which the ineffective rule can be suppressed.

	最大値	最小値
従来法	12000	750
提案手法 1	800	200
提案手法 2	1998	141

表 2　数値実験での無効ルールを抑制するのに必要なエピソード数の計測結果

Table 2　The experimental result of episodes in which the ineffective rule can be suppressed.

	最大値	最小値	平均値	標準偏差
従来法	2766	1181	1581	429.0
提案手法 1	761	547	623	61.8
提案手法 2	1049	525	674	140.3

4.2 迷 路 問 題

4.2.1 実 験 方 法

図 7 に示す変化する迷路を解く問題を対象とする. 学習器はスタートから出発し, ゴールにたどり着いた時点で報酬を得る. 学習器は上下左右の 4 方向に 1 コマ移動することができる. 壁があるマスに移動することは許さない. 一回行動を試みる毎に 1 ステップ経過

したと呼ぶ. 各ルールの初期値は 1.0×10^8, 報酬値は 1.0×10^9, 強化関数は公比 5 の等比減少関数とし, ルールの選択は重みに比例した確率に基づくルーレット選択を採用した. キューの長さを 3.0×10^6 (但し,1 エピソード毎の更新は必要とするメモリ量が膨大となるため, 3000 エピソード毎に更新される長さ 1000 のキューで実装した), 忘却率を 0.9999999 とする. 4.0×10^6 エピソード経過後に環境変化を生じさせ, 乱数の種を変えて各々30 回ずつ計測した.

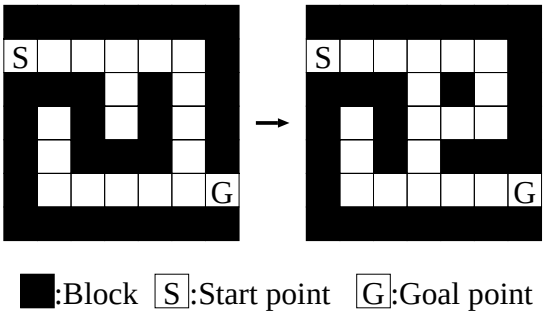


図 7 迷路問題の実験環境; 左:環境変化前, 右:環境変化後
Fig. 7 Grid world for this experiment

4.2.2 実験結果および考察

図 8 に環境が変化しない場合での学習曲線, 図 9 に環境変化後における学習曲線を示す. 報酬までのステップ数が低ければ低いほど, 環境に適応していると考えることができる.

図 8 から, 提案手法が環境が変化しない場合でも従来法と同程度の性能を出すことができることが分かる. また, 環境が変化した場合はより高速に環境に適応していくことが図 9 から分かり, 無効ルールを抑制することにより性能の向上に貢献するということが確認できる.

4.3 追跡問題

4.3.1 実験方法

図 10 に示す 9×9 格子状トーラス環境を設定し, そこに 4 つのハンターエージェントと獲物エージェントをランダムに配置したものを初期状態とする. 各エージェントは上下左右の方向に 1 マス進むかまたは停止の行動をする. 同一のマスに複数のエージェントが存在することは許さない.

ハンターの視界は 5×5 で, 自分の周囲 $5^2 - 1$ マスを見ることができる. また, 自分以外の個々のハンターの区別はしない. 図 10 のように, 全てのハンターが獲物に隣接した状態を目標状態とし, 全てのハンターに

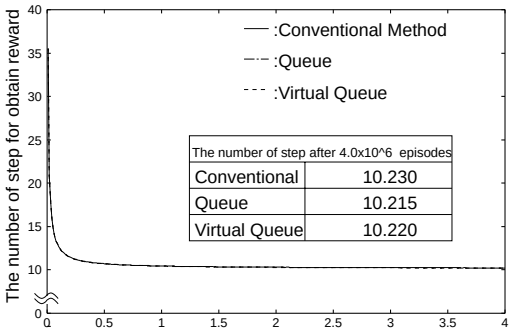


図 8 迷路問題の環境変化前の学習曲線
Fig. 8 Learning result on static environment

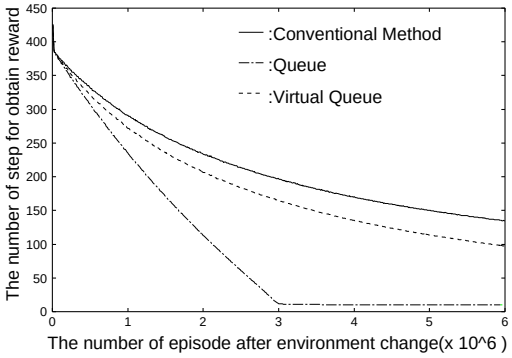


図 9 迷路問題の環境変化後の学習曲線
Fig. 9 Learning result after environment change

報酬が与えられる.

獲物は, ハンターの位置とは無関係にランダムに行動を選択する. 獲物は学習しない.

各ハンターは, 通信や情報の共有無しに, 報酬を唯一の手掛かりとして, それぞれ独立に学習する.1 エピソード毎に各ハンターは学習するので, ハンター個体から見た場合, 環境が 1 エピソード毎に変化していることになる.

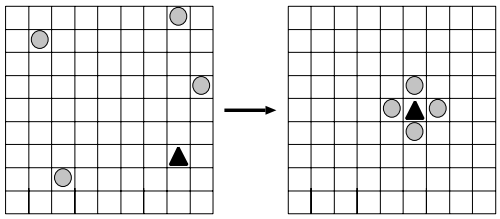


図 10 追跡問題
Fig. 10 Pursuit problem

ハンターの各ルールの初期重みを 1.0×10^5, 報酬値を 1.0×10^6, 公比 5 の等比減少関数を強化関数とする. 忘却率を 0.99999999 とし, キューは 2.5×10^6 エピソード後に更新される長さ 10 のキューで実装される. また, 各手法ともに乱数の種を変えて 40 回ずつ計測し, 平均値を実験結果とする.

4.3.2 実験結果および考察

学習回数と報酬までのステップ数の関係を図 11 に示す. 提案手法 1 はキューが溢れた時点から環境へ適応することができなくなった. これは, キューを更新する毎に 2.5×10^6 エピソードだけの学習成果を捨ててしまう為であると考えられ, キューを更新するエピソード数を短くすればこの問題に対しても適応できると予測できる. 提案手法 2 は徐々に変化していく環境に対しても有効であることが確認できる.

実験結果から, 学習器のメモリが十分でない場合は, 提案手法 1 はその性能を発揮できないことが分かる. 一方, 提案手法 2 は従来手法と必要なメモリの量が同一なので, 従来手法を用いることができる問題に対して適応させることが可能である.

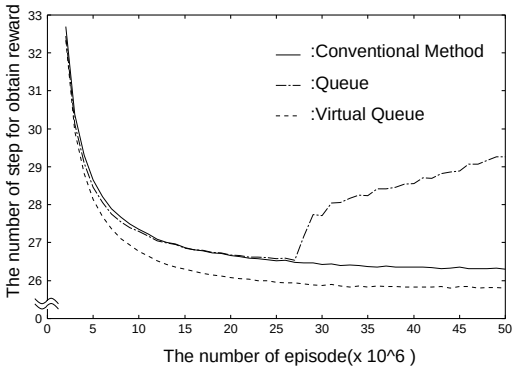


図 11 追跡問題での学習曲線
Fig. 11 Learning result of pursuit problem

5. ま と め

本研究では,Profit Sharing 学習法における環境変化が生じた場合における, 無効ルール抑制に必要なエピソード数について理論的考察を行い, 変化前のエピソード数と抑制に必要なエピソード数について明らかにした. また環境変化後に一定エピソード以内に無効ルールを抑制することが可能な忘却型 Profit Sharing を提案し, 実験によりその有効性を確かめた.

今回提案した手法は, どの程度のエピソードで学習が収束するかがあらかじめ分かっている場合には非常に有効であり, 実験においても忘却率等は従来手法の実験結果を元に, 忘却率などのパラメタを設定した. しかし, 忘却率の変化が学習にどのような影響を及ぼすかについての調査はまだ不十分であり, 今後最適な忘却率の設定方法について検討していきたい. また, 学習の収束についての前提知識を持っていることは一般的であるとはいえない. そこで, そのような前提知識を得られない場合への適用についての検討も今後の課題である.

文 献

[1] Leslie Pack Kealbling, Michael L. Littman, and Andrew W. Moore: Reinforcement Learning: A Survey, Journal of Artificial Intelligence Research 4, pp237-285(1996)

[2] 山村 雅幸, 宮崎 和光, 小林 重信: エージェントの学習, 人工知能学会誌, Vol.10, No.5, pp23-29(1995)

[3] 宮崎 和光, 山村 雅幸, 小林 重信: “MarcoPolo:報酬獲得と環境同定のトレードオフを考慮した強化学習システム” 人工知能学会誌, Vol.12, No.1, pp.78-88(1997)

[4] 片山 晋, 武市 正人, 小林重信: 満足化原理に基づく強化学習のための確率的戦略, 人工知能学会誌, Vol.13, No.6, pp.971-980(1998)

[5] 宮崎 和光, 小林 重信: 離散マルコフ決定下での強化学習, 人工知能学会誌, Vol.12, No.6, pp.811-821(1997)

[6] 木村 元, Lesile Pack Kaelbling: 部分観測マルコフ決定過程での強化学習, 人工知能学会誌, Vol.11, No.4, 1996

[7] 宮崎 和光, 荒井 幸代, 小林 重信: POMDPs 環境下での決定的政策の学習, 人工知能学会誌, Vol.14, No.1, pp.148-156(1999)

[8] 浅田 稔: 強化学習の実ロボットへの応用とその課題, 人工知能学会誌, Vol.12, No.6, pp.831-836(1997)

[9] 木村 元, 小林 重信: 確率的傾斜法を用いた強化学習とロボットへの適用, 電学論, Vol.119-C, No.8(1999)

[10] Jose del R. Millan: Incremental Acquisition of Local Networks for the Control of Autonomous Robots, 7th International Conference on Artificial Neural Networks, pp. 739-744. Special Session on ”Adaptive Autonomous Agents”. Lausanne, Switzerland. Invited paper(1997).

[11] 山口 智浩, 三浦 正宏, 谷内田 正彦: 適応型模倣による複数個体の強化学習, 人工知能学会誌, Vol.12, No.2, pp.323-331(1997)

[12] 荒井 幸代, 宮崎 和光, 小林 重信: マルチエージェント強化学習の方法論–Q-learning と Profit Sharing による接近–, 人工知能学会誌, Vol.13, No.4, pp.609-617(1998)

[13] 宮崎 和光, 荒井 幸代, 小林 重信: Profit Sharing を用いたマルチエージェント強化学習における報酬配分の理論的考察, 人工知能学会誌, Vol.14, No.6, pp.1156-1164(1999)

[14] Robert H. Crites and Andrew G. Barto: Improving Elevator Performance Using Reinforcement Learning, In D. Touretzky, M. Mozer, and M. Hasselmo, editors, Neural Information Processing Systems 8, 1996

[15] 木村 元, 山村 雅幸, 小林重信: “部分観測マルコフ決定過

程下での強化学習:確率的傾斜法による接近", 人工知能学会誌, Vol.11, No.5, pp.761-768(1996)
[16] 山本 真也, 山口 文彦, 斎藤 博昭, 中西 正和: 強化学習における環境変化認識法, 信学技報, AI99-81,pp.31-36(2000-01)
[17] Devika Subramanian, Peter Druschel, and Johnny Chen: Ants and reinforcement learning: A case study in routing in dynamic networks, In Proceedings of IJCAI-97, (1997)
[18] Justin A. Boyan and Michael L. Littman.: "Packet routing in dynamically changing networks: A reinforcement learning approach", In Jack D. Cowan, Gerald Tesauro, and Joshua Alspector, editors, Advances in Neural Information Processing Systems, volume 6, pages 671–678. Morgan Kaufmann, San Francisco CA, (1993)
[19] 宮崎 和光, 山村 雅幸, 小林 重信: 強化学習における報酬割り当ての理論的考察, 人工知能学会誌, Vol.9, No.4, pp.580-587(1994)

付　　録

1. 補題 1 の証明

環境に変化が生じない場合, 唯一の回帰的無効ルールを抑制することが最も困難であることは宮崎ら [19] により明らかにされている.

変化後の環境に注目した場合, 変化前の環境は各ルールの初期値に対してのみ影響を与えるといえる. また, 無効ルールの初期値が有効ルールの初期値と比較して大きいほど抑制が困難になるのは明らかであり, 変化前の環境で得られる報酬を独占した場合に他のルールとの差が最大となる.

ゆえに, 環境変化後の唯一の回帰的無効ルールが変化前の同一感覚入力下で得られる報酬を独占していた場合が無効ルールを抑制するのに最も困難な条件である.

(証明終り)

2. 補題 2 の証明

まず, この証明で用いる変数の定義をする. 環境の変化前に有効ルールが得られる報酬を Rw', 変化後に得られる報酬を Rw, 変化前の環境のエピソード数を G, 変化後のエピソード数を X, 有効ルールの数を L, 式 (4) に示した無効ルールが得られる最大の比率を P で表す. また, 無効ルールが環境変化前に得た報酬を $W'(0)$, 変化後に得た報酬を $W(0)$, 任意の有効ルールが変化前に得た報酬を $W'(i), (i \neq 0)$, 変化後に得た報酬を $W(i)$ で表す.

次式を満たしているとき無効ルールは抑制されている.

$$W'(0) + W(0) < W'(i) + W(i) \qquad \text{(A·1)}$$

まず, 変化前の環境で得た報酬について考えると

$$\begin{cases} W'(0) &= G \cdot Rw' \\ W'(i) &= 0 \end{cases} \qquad \text{(A·2)}$$

であるのは明らかである. また, 変化後の報酬について有効ルールが得る報酬と無効ルールが得る報酬の差が最小になる条件で考えた場合は,

$$\begin{cases} W(i) &= Rw \cdot \frac{X}{L} \\ W(0) &= P \cdot W(i) \end{cases} \qquad \text{(A·3)}$$

によって表される. ゆえに, 式 (A·1) を展開して,X について解くと

$$G \cdot Rw' + P \cdot Rw \cdot \frac{X}{L} < Rw \cdot \frac{X}{L}$$

$$\frac{L \cdot Rw'}{(1-P)Rw} \cdot G < X \qquad \text{(A·4)}$$

(証明終り)

3. 定理 2 の証明

定理 1 の証明に用いたように, 最も困難条件下での無効ルールの抑制に必要なエピソード数について考える.

補題 2 で用いた変数に加えて, キューの長さを Q_{length} で表す. また,$G + X \leq Q_{length}$ の場合は, 無効ルールの抑制に必要なエピソード数について従来手法と変わらないので, ここでは $G + X > Q_{length}$ の場合について考える. 環境変化後に得る報酬は従来法と同様に

$$\begin{cases} W(i) &= Rw \cdot \frac{X}{L} \\ W(0) &= P \cdot W(i) \end{cases} \qquad \text{(A·5)}$$

で表せる. 一方,Q_{length} を超えて報酬を保持することはできないので, 変化前に得た報酬は

$$\begin{cases} W'(0) &= (Q_{length} - X) \cdot Rw' \\ W'(i) &= 0 \end{cases} \qquad \text{(A·6)}$$

となる. ゆえに, 式 (A·1) を展開すると,

$$(Q_{length} - X) \cdot Rw' + P \cdot Rw \cdot \frac{X}{L} < Rw \cdot \frac{X}{L} \qquad \text{(A·7)}$$

$$\frac{Q_{length} \cdot L \cdot Rw'}{(1-P)Rw + L \cdot Rw'} < X \qquad \text{(A·8)}$$

となり, 最も困難な条件下で無効ルールを抑制する為の条件は式 (A·8) で表される. 推移律により, 任意の条件下で無効ルールを抑制するために必要な条件は式 (A·8) で表される.

(証明終り)

4. 定理 3 の証明

定理 1 の証明に用いたように, 最も困難条件下での無効ルールの抑制に必要なエピソード数について考える.

この証明で用いる変数として, 補題 2 で用いた変数に加えて忘却率を τ で表す.

まず, 変化前の環境で得た報酬は式 (10) より,

$$\begin{cases} W'(0) &= \sum_{j=X+1}^{G+X} Rw' \cdot \tau^{j-1} \\ W'(i) &= 0 \end{cases} \quad \text{(A·9)}$$

で表せる. また, 有効ルールに与えられる報酬が L 等分されるとき無効ルールの抑制が最も困難になるので, 変化後に得られる報酬は

$$\begin{cases} W(0) &= P \cdot W(i) \\ W(i) &= \frac{1}{L} \sum_{j=1}^{X} Rw \cdot \tau^{j-1} \end{cases} \quad \text{(A·10)}$$

となる. ゆえに, 式 (A·1) を展開すると,

$$\sum_{j=X+1}^{G+X} Rw' \cdot \tau^{j-1} + \frac{P}{L} \sum_{j=1}^{X} Rw \cdot \tau^{j-1} < \frac{1}{L} \sum_{j=1}^{X} Rw \cdot \tau^{j-1} \quad \text{(A·11)}$$

$$\frac{Rw' \cdot \tau^X \cdot (1-\tau^G)}{1-\tau} < \frac{Rw(1-P)(1-\tau^X)}{L(1-\tau)} \quad \text{(A·12)}$$

$$\tau^X < \frac{(1-P)Rw}{(1-P)Rw + L \cdot Rw' \cdot (1-\tau^G)} \quad \text{(A·13)}$$

となり, 最も困難な条件下で無効ルールを抑制するのに必要なエピソード数について条件が得られる. 推移律により式 (A·13) を満たせば, 環境変化後の任意の無効ルールを抑制できる.

(証明終り)

(平成 x 年 xx 月 xx 日受付)

加藤　新吾　(学生員)

1999 名古屋工業大学電気情報工学科卒業. 1999 同大学院修士課程, 現在に至る.

松尾　啓志　(正員)

昭和 58 年 名古屋工業大学情報工学科卒業. 昭和 60 年 同大学院修士課程修了. 同年松下電器産業入社. 昭和 61 年 同大学院博士前期課程入学. 平成元年 同大学院博士課程修了. 平成元年 名古屋工業大学電気情報工学科助手. 平成 7 年 名古屋工業大学電気情報工学科助教授, 現在に至る. 分散システム, 分散処理, パターン認識に関する研究に従事.