特　集

Profit Sharing に基づく強化学習の理論と応用

Theory and Applications of Reinforcement Learning based on Profit Sharing

宮崎　和光*　　木村　元*　　小林　重信*
Kazuteru Miyazaki　Hajime Kimura　Shigenobu Kobayashi

* 東京工業大学大学院総合理工学研究科
Graduate School of Interdisciplinary Science and Engineering, Tokyo Institute of Technology, Yokohama 226-8502, Japan.

19YY 年 MM 月 DD 日　受理

Keywords: Reinforcement Learning, Profit Sharing, Multi-agent System, Rationality Theorem, Rational Policy Making.

1. はじめに

1・1　工学の視点からみた強化学習

強化学習とは, 報酬という特別な入力を手がかりに環境に適応した行動決定戦略を追求する機械学習システムである. 強化学習の重要な特徴に, 1) 報酬駆動型学習であること, 2) 環境に対する先見的知識を前提としないこと, 2 点がある. このことは, 「何をして欲しいか (what)」という目標を報酬に反映させるだけで, 「その実現方法 (how to)」を学習システムに獲得させることを意味する.

強化学習システムは, 人間が考えた以上の解を発見する可能性がある. 加えて, 環境の一部が予め既知な場合には, 知識を組み込むことも可能である. この場合, 知識ベースが不完全であってもあるいは多少の誤りが含まれていても構わない.

また, 強化学習は, ニューロやファジイなどの既存の手法との親和性が高い. さらに, 緩やかな環境変化には追従可能である. これらの理由から, 強化学習は工学的応用の観点から非常に魅力的な枠組と言える.

1・2　強化学習の要件

上記の特徴に加え, 著者らはさらに, 以下のふたつの要件が, 強化学習を工学に応用する際, 重要であると考える.

要件 1　適用可能な環境のクラスが広いこと.

要件 2　素早い学習が可能であること.

要件 1 については, 実際に, 強化学習で「何がどの程度実現可能か」という問題に関係する. 適用可能なクラスは広いほど好ましいが, 期待獲得報酬量を最大化するという意味での**最適性**を要求する場合, 適用可能な環境のクラスは限定される. したがって, 適用可能なクラスを広げようとすると, 最適性を犠牲にしなければならない. この場合, 何らかの合理性が保証されることが望ましい. このように, 環境のクラスと解の質の間にはトレードオフの関係が存在する.

要件 2 については, 学習初期に試行錯誤による報酬獲得を誘導し, かつ得られた報酬を適切にフィードバックする必要がある. 後者は, 学習アルゴリズムの工夫により実現可能であるが, 前者は, 報酬をどのように設計するかという問題に関係する.

報酬設計で, 最も明快な方法は, 目標を達成したときのみ報酬を与えることである. しかし, この場合, 学習アルゴリズムを工夫したとしても, 問題によっては, 最初の報酬を得るまでに, 非常に多くの試行錯誤的行動を要する心配がある. そこで, 副目標に対する報酬や制約違反に対する罰を導入することが考えられる. しかし, 図 1に示すように, 副目標の導入が副作用を及ぼす場合もあるので, 報酬や罰の設計は慎重にしなければならない.

強化学習の研究は**環境同定型**と**経験強化型**に類別される [山村 95]. 現在,Dynamic Programing (DP) に基づく環境同定型が主流とされているが, この接近法は, 上記の要件 1 および 2 を十分に満たしているとは言えない状況にある. 一方, 経験強化型の Profit Shar-

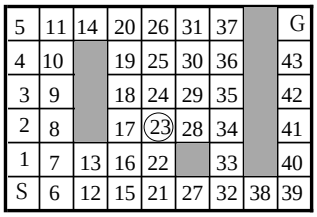

図 1 報酬設計の困難さを示すための例. 始点 S から終点 G への学習を加速させるために, それらのほぼ中間地点である状態 23 を副目標に設定した場合を考える. このとき, 状態 23 に与える報酬値によっては, G に到達しないで S と状態 23 の間を往復したり, あるいは, わざわざ状態 23 を経由して G に到達する解が学習される可能性がある.

ing (PS)[Grefenstette 88] は, 上記ふたつの要件を満たす手法として有望である.

1・3 用語の定義

以降の議論において必要となる用語を定義する. 本稿で対象とする強化学習システムは, 環境からの**感覚入力**に対し, **行動**を選択し, 実行に移す. 一連の行動に対し, 環境から**報酬**が与えられる. 時間は認識-行動サイクルを 1 単位として離散化される. 感覚入力は離散的な属性-値ベクトルである. 行動は離散的なバリエーションから選ばれる.

ある感覚入力において実行可能な行動はルールとして記述される. 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 xa と書く. 初期状態あるいは報酬を得た直後から次の報酬までのルール系列を**エピソード**という. 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**という. 現在までのすべてのエピソードで, つねに迂回系列上にあるルールを**無効ルール**と呼び, それ以外を**有効ルール**と呼ぶ.

各感覚入力に対し, 選択すべき行動を与える関数を**政策**と呼ぶ. 単位行動当たりの期待獲得報酬が正である政策を**合理的政策**と呼び, それを最大化する政策を**最適政策**と呼ぶ. さらに, 各感覚入力に対し, 選択すべき行動をひとつだけ与える関数を**決定的政策**いくつかの行動を確率的に与える関数を**確率的政策**と呼ぶ.

本稿では, PS に基づく強化学習の理論と応用に焦点をあて解説する. 以下, 第 2 章では,DP に基づく接近の特徴と限界を論じる. 第 3 章では,PS に基づく接近について, PS の合理性定理を紹介した後, PS と適性度の履歴との関係, 決定的政策に特化した学習について述べる. 第 4 章では,PS の具体的な適用事例を紹介することにより, PS の工学的応用の可能性を示す. 第 5 章は, まとめであり, 今後の課題をとりまとめる.

2. DP に基づく強化学習

2・1 MDPs を対象とする強化学習

機械学習における強化学習は [Samuel 59] の checker player に端を発するが, [Sutton 88] の Temporal Difference 法 (TD) や [Watkins 92] の Q-learning(QL) における理論的進展により, 近年, 強化学習への関心が高まっている.

TD や QL は, 強化学習の対象問題を DP の問題として捉えることにより, 最適性の議論を可能にしている. 特に QL は, Markov Decision Processes (MDPs) の環境下で割引期待獲得報酬を最大化するという意味での最適性が保証される. 一方,MDPs の環境が既知な場合, Policy Iteration Algorithm(PIA)[ワグナー 78] により最適性が保証される.

ある未知環境が MDPs ならば,QL により最適性が保証されるという点は, 理論的および工学的な観点からも非常に重要である. すなわち, もし MDPs 環境下で最適性が要求されるときは, QL や k-確実探査法 [宮崎 95] および MarcoPolo[宮崎 97a] などの PIA に基づく手法を用いればよい. しかし, 実問題がつねに MDPs で記述可能であるとは限らない. その意味で,DP に基づく手法は, 要件 1 を満たしているとは言えない.

QL 等の学習は, 報酬が徐々にその周辺の状態に伝搬していく形で進行する. 例えば, 図 1の環境では, G で報酬が得られた後, 状態 43 が学習され, 状態 43 の学習が成功した後はじめて, 状態 42 の学習が成功する. さらに, 最適性を保証するためには, この他に, 与えられた環境を同定するための行動が必要になる. このような理由から QL に代表される DP に基づく強化学習手法は学習が非常に遅く, 要件 2 を満たしているとは言えない.

2・2 MDPs を超えるクラスを対象とする強化学習

現在,MDPs を超えるクラスとして, Semi-Markov Decision Processes (SMDPs)[Bradtke 94] や Partially Observable Markov Decision Processes (POMDPs) [Singh 94] に関する研究が盛んである [木村 97, 木村 99].

SMDPs は時間が連続な MDPs であり, 報酬が時間積分される点を除けば, 基本的な取り扱いは MDPs と大きな相違はない. 現状では,SMDPs は,QL を変形して解かれている [Bradtke 94]. そのため, 要件 2 に関

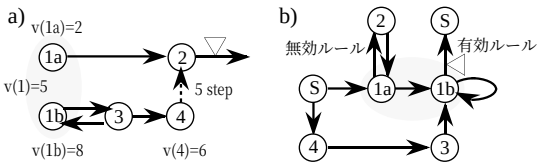

図 2 状態 1a と 1b を混同した結果生じる, a) タイプ 1 の混同の例. b) タイプ 2 の混同の例. [宮崎 99a].

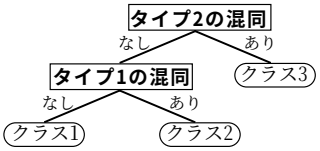

図 3 タイプ 1, タイプ 2 の混同による環境の分類.

し,QL 等と同様の問題を有する.

一方,POMDPs とは, 実際には異なる環境の状態が学習器にとっては同一の感覚入力として知覚される, いわゆる**不完全知覚状態** [Whitehead 91] を有する問題クラスのことをいう. このクラスに対する最も伝統的な接近法は, 過去の履歴を用いて不完全知覚状態を分離する**メモリーベース法** [Chrisman 92, McCallum 95] である. そこでは不完全知覚状態を分離した後の学習には, 通常 QL が用いられる. そのため要件 2 に関し,QL 等と同様の問題を有する. また, メモリーベース法は, 一般に非常に多くのメモリーを要すること, および過去の履歴の収集に膨大な試行錯誤を必要とすることなどが問題点として挙げられる.

現在, メモリーベース法の欠点を克服するために**確率的政策** [Singh 94, Jaakkola 94, 木村 96] が提案されている. そこでは確率的に行動を選択することにより, 不完全知覚状態からの脱出を試みる. 確率的政策は, 決定的政策により合理性が保証されないクラスに対しては有効であるが, それが保証されるクラスでは, 報酬を得るために必要以上に多くの行動を要してしまう場合があり, 必ずしも有効とはいえない.

2・3 MDPs を超えるクラスの困難さ

著者らは, 強化学習が取り扱う環境の困難さを, **タイプ 1** および**タイプ 2 の混同**という二つの観点から分類している [宮崎 99a]. また, この考えは, 各エージェントが独立に学習するマルチエージェント強化学習に対してもそのまま適用可能である [荒井 98, 宮崎 99b].

図 2a) に示すように, 価値の高い状態と価値の低い状態が同一視されることをタイプ 1 の混同と呼ぶ. 図 2a) において, 状態の価値を報酬への最短ステップ数で見積もる場合, 状態 1a の価値は 2, 状態 1b の価値は 8 となる. 状態 1a と 1b は実際には異なる状態であるが, 学習器には同一の状態 (状態 1) として知覚される. したがって, 状態 1a と 1b を等確率で経験したとすれば, 状態 1 の価値の期待値は 5 となり, 状態 4 の価値である 6 よりも高くなる. その結果, 状態 3 では**左**すなわち状態 1b へ向かう行動が最適とされ, 状態 1b と 3 の間を往復する非合理的な政策が学習される.

また, 図 2b) に示すように, あるとき有効ルールであると判定されたルールが, ある時点以降, つねに迂回系列上に存在してしまうことをタイプ 2 の混同と呼ぶ. 図 2b) において, 状態 1b で**上**という行動は有効ルールであるが, 同じ行動は状態 1a ではつねに迂回系列上に存在する. 学習器は状態 1a と 1b をともに同じ状態 (状態 1) と認識するため, 状態 1 で**上**という行動は, 学習器にとっては有効ルールとされる. しかし, たとえそのルールを選んだとしても, 状態 S で**右**へ向かう行動を学習した場合には, 状態 1a と 2 の間を往復する非合理的な政策が学習される.

一般に, タイプ 2 の混同が存在すれば, タイプ 1 の混同も同時に存在する. タイプ 1 およびタイプ 2 の混同という観点から, 環境は図 3に示すような 3 つのクラスに分類される. MDPs はクラス 1 に属する. QL に代表される DP に基づく手法は, 状態の価値を推定することにより学習が進行するので, タイプ 1 の混同の影響を強く受ける. そのため, クラス 1 以外では, 非合理的な政策を学習する可能性が高い. 工学的応用の観点からはクラス 2 や 3 に対しても頑健な手法が望まれる.

3. PS に基づく強化学習

3・1 PS の合理性定理

1.2 節で述べたように, Profit Sharing (PS) は強化学習のふたつの要件を満たしている. PS とは, 報酬を得たときに, それまでに使用されたルール系列を, 一括的に強化する手法である. PS ではエピソード単位でルールに付加された評価値を強化する. 報酬からどれだけ過去かを引き数とし, 強化値を返す関数を**強化関数**と呼ぶ. 時点は離散なので f_iによって報酬から i ステップ前の強化値を参照する. 長さ lのエピソード $(r_l \cdots r_i \cdots r_2 \cdot r_1)$ に対して, ルール r_iの重みである ω_{r_i}は, $\omega_{r_i} = \omega_{r_i} + f_i$によって強化される.

[宮崎 94, 宮崎 99] によれば, タイプ 2 の混同が存在しないクラスにおいて, 以下の定理が成立する.

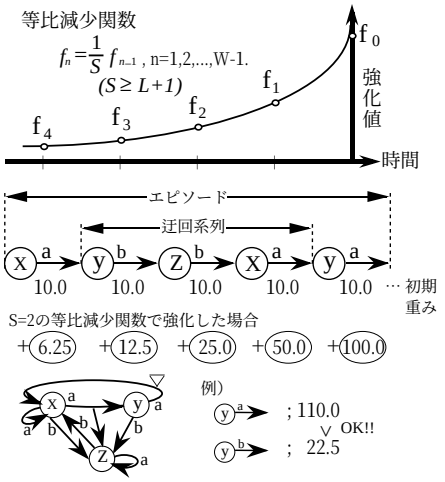

図 4　定理を満たす等比減少関数の例

［定理 1］（PS の合理性定理）
タイプ 2 の混同が存在しない環境下で, 合理性を保証するための強化関数の必要十分条件は

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

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. □

一般に,L の値は学習以前には知ることができないが, 実装にあたっては, L を可能な行動出力の種類引く 1 とすれば十分である. 以下では, この条件を**無効ルール抑制条件**と呼ぶ. 定理を満たす最も簡単な強化関数としては, 図 4に示す等比減少関数が考えられる.

PS は状態の価値を学習に利用しないので, タイプ 1 の混同の影響を一切受けない. また, タイプ 2 の混同が存在しなければ, MDPs を超えるクラスにおいても, つねに合理的政策の獲得が保証される (図 3参照). また, エピソード単位でルールを強化するため, 一度の報酬で多くのルールを強化することができ, 学習効率がよい. 報酬の値に関しては,QL 等と同様, 複数の報酬値を設定することも可能だが, PS は,1 種類の目標に対し 1 種類の正の報酬を与える環境下での学習に最も適している.

3・2　PS と適性度の履歴との関係

PS はエピソードという形で履歴情報を学習に利用している. 類似の概念に**適性度の履歴** (eligibility trace) [Sutton 98] がある. PS は報酬を得たときに過去の履歴をまとめて更新するのに対し, 適性度の履歴は, 報酬の有無に関わらず, ステップ毎, 過去の履歴を参照して更新される.

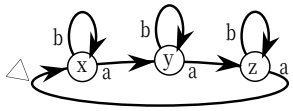

図 5　PS と適性度の履歴の関係を議論するために利用した環境.

履歴情報を学習に利用する際に, 例えば図 5に示す環境では, 無効ルール xb,yb,zb の抑制が必要であるが, 適性度の履歴を用いた場合, これらのルールが抑制されるとは限らない [Sutton 98]. 一般に, 適性度の履歴を用いた場合に, このようなルールを抑制するためには,Replacing Trace という工夫が必要になる [Sutton 98].

適性度の履歴には, 過去の履歴情報の重要度を制御するパラメータ λ が存在する. このパラメータは PS の強化関数に類似するものであり, λ に関し,PS の合理性定理と同様の定理が存在する可能性がある. そのような定理が見い出されれば, PS と適性度の履歴との関係はより明確になる.

また, 適性度の履歴は, 一般に,TD や QL と組み合わせて利用される. この意味から, 逆に PS の TD や QL との組み合わせも考えられる. 具体的研究事例として,[堀内 99] が存在するが, その他の組合せ方法も今後の課題として興味深い.

3・3　POMDPs 下での決定的政策の学習

有効ルールの定義より, あるエピソードにおいて同一の感覚入力に対する行動選択の中で報酬に最も近い位置で選択された行動を含むルールは有効ルールである. この性質を利用し, 合理的政策のより効率的な獲得を目指した手法として Rational Policy Making algorithm (RPM) [宮崎 99a] がある (図 6参照).

RPM は決定的な合理的政策が存在するクラスにおいては, タイプ 2 の混同の有無に関わらず, PS よりもさらに効率よく合理性を保証することが可能である. ただし, 現在は, 政策の更新は**マルチスタート法**と呼ばれる政策を最初から形成し直す手法に依存しており, 効率が悪い. マルチスタート法への依存を緩和した手法の開発は, 現在, 研究中である [宮崎 98].

RPM の利用方法としては, 1) 決定的な合理的政策が存在する領域は RPM で学習させ, それ以外の領域は PS で学習させる方法や, 2)RPM と PS を同時に利用し, RPM で政策が得られない場合に限り, PS の学習結果を利用する方法, 等のハイブリッド的な手法が

procedure Rational Policy Making algorithm (RPM)
begin
  do
    1次および2次記憶領域の内容を初期化する.
    do
      if 現在の感覚入力の2次記憶上に行動が
        記憶されている then その行動を出力する.
      else 環境探査戦略による行動を出力し,
        その行動を1次記憶上に上書きする.
      if 報酬を得た then 1次記憶領域の内容を
        2次記憶領域に複写する.
    while (2次記憶領域が未収束**)**
    if 合理的政策が得られている then
      2次記憶領域の内容を保存する.
  while
end.

図 6　RPM のアルゴリズム. [宮崎 99a].

有望である.

4. PS に基づく強化学習の適用事例

PS に基づく強化学習の適用事例として 4 つの事例を紹介する. 最初のふたつはシングルエージェント系での適用事例であり, 後のふたつはマルチエージェント系での適用事例である.

4・1　The acrobot problem への適用

Single-agent 系での具体的な適用事例として The acrobot problem[Sutton 98] を紹介する.

〔1〕 The acrobot problem

The acrobot problem とは図 7に示すような 2 リンク (リンク長 $\ell_1 = \ell_2 = \ell$) からなるアームの先端を第 1 関節の上方 ℓ以上に振り上げる問題である. 状態変数は各関節の角度 (θ_1,θ_2) および角速度 $(\dot{\theta_1},\dot{\theta_2})$, 行動は第 2 関節に加えるトルク$\tau = \{+1,0,-1\}$ である. 初期状態は$\theta_1 = \theta_2 = \dot{\theta_1} = \dot{\theta_2} = 0.0$, すなわち, リンクがまっすぐ下にぶら下がった状態である.

連続値である角度を i $(i = 2, 6, 12)$ 等分, 角速度を $1.5i$ 等分し, それぞれの分割の組 (角度-角速度 : 2-3, 6-9, 12-18) に対し,PS,SGA,RPM で実験を行った. なお, 変数の離散化に伴い, この問題は, 図 3のクラス 2 の POMDPs に属するのでタイプ 1 の混同が存在する.

〔2〕 結果および考察

各手法において, 乱数の種を変えて行った 100 回の実験の結果を図 8に示す. 横軸は行動選択回数, 縦軸はその時点で得られた政策による報酬獲得までに要した平均行動数である.

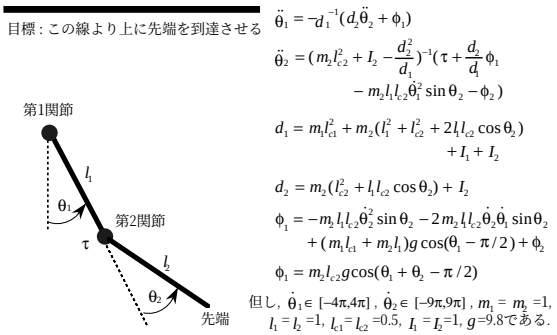

図 7　The acrobot problem.

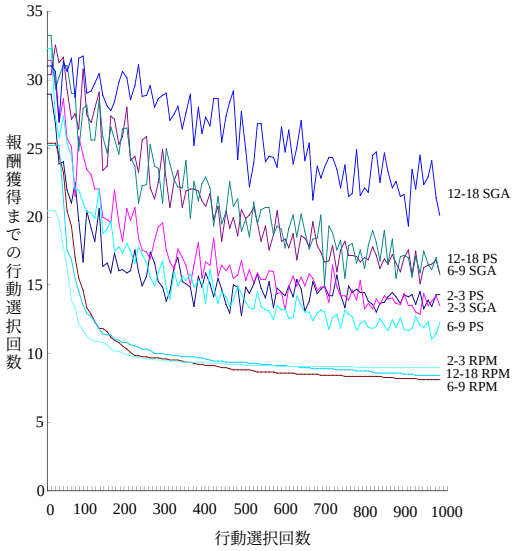

図 8　The acrobot problem への適用結果.

本問題には, 決定的な合理的政策が存在するにも関わらず, SGA では, 確率的政策が獲得され, 他手法に比べ性能が悪化している. 特に分割が細かい場合この傾向が顕著に現われる. PS は SGA よりも確率的政策が得られる度合は低く,SGA より性能は良い.

この実験例では, 決定的な合理的政策が存在するので, RPM は効率良く政策を学習するのに成功している. ただし分割が粗い場合, 例えば, 2-3 分割で, トルクを$\frac{1}{10}$にし, 原点付近で初期化した場合, 決定的な合理的政策が存在しなくなるので,RPM は適用できない. この場合,RPM は,PS や SGA と組み合わせるハイブリッド的な手法が有効である.

4・2　ロボットへの適用

ここでは, ロボットへの適用の一例として Lego 社の

図 9　Lego ロボット.

$MindStorms^{TM}$を利用した事例を紹介する.

〔1〕 Lego ロボットへの強化学習の実装

図 9に示す Lego ロボットを利用した. RPM で行動コマンドレベル (Move,Turn,Hand) を学習した後に PS で餌を得るためのルール (感覚-行動対) の学習に入るハイブリッド法を実装した. プログラミング言語には,NQC[古川 99] を利用した.

RPM による行動コマンドレベルの学習は, 出力ポート A,B,C の選択, および選んだポートへの出力 (+1 または-1) の組合せが学習の対象となる. ポート A と C を選択し,A に+1,C に-1 を出力した場合 Move, A と C を選択し,A に+1,C に+1 を出力した場合 Turn, B を選択し,+1 を出力した場合 Hand コマンドがそれぞれ成功する.

PS の感覚入力は, 光センサー (LS) および接触センサー (TS) の情報を基に決定される. 具体的には, LS と TS がともに off のとき「何も見えない」, LS のみが on のとき「餌を発見」, LS と TS がともに on のとき「餌に接触」という感覚入力を得る. PS の学習は,「餌の発見」,「餌への接近」,「餌に接触し,Hand コマンドを出力する」の 3 段階を要する. この問題は, 図 3のクラス 3 の POMDPs に属するのでタイプ 1 とタイプ 2 の混同が存在する.

〔2〕 動作結果

RPM によるコマンドレベルの学習は直ちに達成される.PS による餌を得るための学習は「何も見えない」状態以外は, [宮崎 97b] で計算機シミュレーションした結果と同一であったが,「何もみえない」状態では [宮崎 97b] と異なり, Turn と Move がほぼ均等に強化されていた. これは実環境では, 環境やロボット自身の動作による不確実性が高く「何もみえない」状態で Turn のみでは餌を視界に捉えるのに不十分であったためと考えられる.

この例のように, 行動コマンドレベルの学習では, 不確実性が存在しないので,RPM によって素早い学習ができる. 目標 (餌の獲得) レベルの学習では, センサーの不完全性に起因する不確実性が存在するが, PS により素早く学習できる. 実ロボットへの適用に際しては, このような 2 レベルの学習が有効である.

表 1　追跡問題における Profit Sharing と Q-learning の比較

		学習方法			
		Q-learning		Profit Sharing	
		平均	分散	平均	分散
環境サイズ	7×7	23.67	7.12	4.75	1.07
	9×9	収束せず		9.68	2.45
	15×15			39.72	12.12

4・3　追跡問題への適用

マルチエージェント系での具体的な適用事例として [荒井 98] の追跡問題への適用例を紹介する.

〔1〕 追跡問題

ここでは追跡問題の典型例として, n×n格子状トーラスの環境に, 2 人のハンターエージェントと 1 匹の獲物エージェントが存在する場合を紹介する. 初期状態では, 各エージェントがランダムに配置され, その後, 予め決められた順番で各エージェントは行動し，上下左右の方向に１コマ進むかまたは停止の行動をひとつ選択する．複数のエージェントが同一場所に存在することは許さない．すべてのハンターが獲物に隣接したときを目標状態とし，すべてのハンターに報酬を与える．その後, ランダムにハンターと獲物を再配置する.

ハンターの視界は５×５とする．獲物の視界は５×５（ただし，斜め方向は死角）とし，視界内にあるハンターから遠ざかる方向に逃げる行動を選択する．獲物は学習しない．ハンターは PS または QL で学習させた．PS では，公比 0.2 の等比減少関数を強化関数に採用した．QL のパラメータは [荒井 98] を参照されたい．この問題は, 図 3のクラス 3 の POMDPs に属するのでタイプ 1 とタイプ 2 の混同が存在する.

〔2〕 結果および考察

環境サイズ n を 7,9,15 とした場合の 10 万エピソード後の報酬までのステップ数の平均と分散を表 1に示す. この問題は, 獲物の逃避的行動に伴う状態遷移の不確定性の増大と不完全知覚の相乗作用により，QL にとっては学習が困難な問題である. 特に, 環境が９×９及び１５×１５の場合，QL は収束せず，政策を形成することに失敗している.

不完全知覚領域の中で多数を占めるのは視界に「何も見えない」状態であり，環境サイズが大きくなるに

つれて，その割合は増加する．そのため「何も見えない」状態での行動の学習は非常に重要である.「何も見えない」状態での QL は 100,000 エピソード付近でも行動選択の割合が振動を繰り返しているのに対し，PS は 5,000 エピソード位で収束する傾向にあり，しかも 2 人のハンターが相補的な行動を強化して，意味のある協調的行動を取るための確率的政策を形成していた. このような挙動の違いが QL と PS の間での性能の差を決定づけていると考えられる.

4・4 クレーン群制御問題への適用

マルチエージェント系でのより複雑かつ実問題を意識した適用事例として [Arai 98] のクレーン群制御問題を紹介する.

〔1〕 クレーン群制御問題

100 番地からなる製鉄所における冷延工場コイルヤードを考える．クレーン群制御問題とは，熱延工場からコイルヤードへと次々に払い出されて来る熱延コイルの受け入れおよび払い出しをヤード内の 3 台のクレーンによって協調的に実行させる問題をいう．以下では, 各熱延コイルの (始点 (運搬元番地)，終点 (運搬先番地)) を与える 2 項組を**タスク**と呼ぶ.

クレーン制御問題は**タスク割り当て** (逐次的に投入されるタスクのクレーンへの割り当て), および，**タスク実行** (割り当てられたタスク実行のためのプリミティヴな行動決定) の 2 つの段階からなる問題解決サイクルとして捉えることができる．タスク割り当てについては，最も近いクレーン優先でトップダウン的に行い, タスク実行のみを強化学習の対象とする．この問題は, 図 3のクラス 3 の POMDPs に属するのでタイプ 1 とタイプ 2 の混同が存在する.

〔2〕 強化学習器の設計

タスク実行のためのルールを獲得するために, 以下のように PS を設計した.

まず, エージェントの行動集合は，目的地方向へ走行する**順走**，および，衝突回避のための**退避**，**待機**の 3 つとする．エージェントの状態集合は，**空走**, **搬走**, **巻き上下**, **遊走**, **待機**, **退避**の 6 つとする (各状態の詳細は [Arai 98] を参照されたい)．クレーン間の衝突は，自クレーンと他クレーンの距離が 3 番地以内になった際, 検出されるので, 感覚入力を自クレーン位置の前後 3 番地に限定した.

クレーンの実行制御におけるすべての衝突回避ルールを強化学習によって獲得させる必要はない．衝突を回避すべき相手および自分の状態がともに空走か搬走の場合を強化学習の対象とする.

評価に用いたタスク集合

タスク	A	B	C	D	E	F	G	H	I	J	K	L
始点	25	43	23	68	44	45	12	80	49	51	41	49
終点	36	25	36	100	27	63	36	100	22	92	36	26

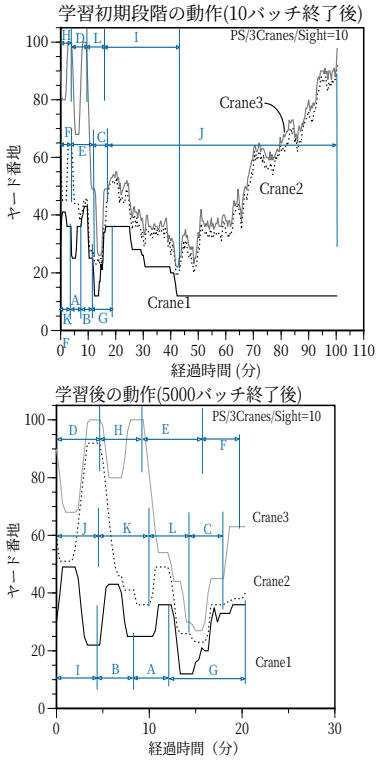

図10 学習初期および学習後の軌跡

報酬は，サブタスク終了ごとに処理をしたエージェントに与えられるものとし，回避などによって貢献した他エージェントには分配しない．現在, 著者らは, マルチエージェント系において, 他エージェントへの報酬配分に関する定理を導出されており [宮崎 99b], その定理を利用した実験は, 現在, 計画中である.

〔3〕 結果および考察

12 タスクからなるタスク集合を 1 バッチとして, 公比 0.5 の等比減少関数を用いた PS で, 5000 回学習を繰り返した．この試行を 5 回繰り返したところ, 1 バッチ終了までの平均所要時間は 33.4 分であった．一方, 学習をせずに「終点までの距離が短い方優先」というトップダウンに設計したルールに基づいたシステムでの所要時間は 40.5 分であった.

さらに，図 10に,PS により得られた学習初期と学習後の軌跡を示す．この図から，学習初期段階では回避行動に無駄が多いが，学習後は，各クレーンの無駄な動きが排除され，適切なクレーン間の衝突回避ルール

が獲得されていることがわかる.

これらの結果より, トップダウンに与えた汎用のルールには限界がある一方, PS を用いた場合には, 各クレーンが担当するタスクの始点と終点の分布の違いに基づいた適切なルールが獲得できていることが確認された.

5. おわりに

強化学習は, 目標達成時に報酬を与えるのみで, 与えられた環境に適応して, 目標達成方法を自動的に獲得する手法である. そのため, 工学的応用の観点からも非常に興味深い枠組である. 強化学習の応用は近年増えつつあるが, まだ多いとは言えない状況にある [木村 99].

本稿では, 強化学習を工学に応用する際, 重要となるふたつの要件を述べ, DP に基づく伝統的接近がそれらの要件を満たさないことを論じた. ふたつの要件を満たす手法として経験強化型の PS の理論と手法を解説した後, 具体的な適用事例を紹介した.

現在, 我々は, マルチエージェント系における報酬配分に関する定理を利用した工学的応用を計画している. 今後の課題としては, 報酬とは異なる軸としての罰の PS における取り扱い, PS に基づく手法のさらなる拡張と改良, 適用可能な問題クラスの拡大, より現実的な問題への応用などが特に重要であると考えている.

◇参考文献◇

[荒井 98] 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. 5, pp.609-618 (1998).

[Arai 98] Arai, S., Miyazaki, K., and Kobayashi, S.: *Cranes Control Using Multi-agent Reinforcement Learning*, International Conference on Intelligent Autonomous System 5, pp.335-342 (1998).

[Bradtke 94] Bradtke, S. J. and Duff, M. O. *Reinforcement Learning Methods for Continuous-Time Markov Decision Problems*, Advances in Neural Information Processing Systems 7, pp.393-400 (1995).

[Chrisman 92] Chrisman, L.: Reinforcement learning with perceptual aliasing: The Perceptual Distinctions Approach, *Proceedings of the 10th National Conference on Artificial Intelligence*, pp. 183-188 (1992).

[Grefenstette 88] Grefenstette, J. J. *Credit Assignment in Rule Discovery Systems Based on Genetic Algorithms*, Machine Learning 3, pp.225-245 (1988).

[古川 99] 古川 剛編, JinSato, 白川裕記, 牧瀬哲郎, 倉林大輔, 衛藤仁郎 共著: *MINDSTORMS* パーフェクトガイド, 翔泳社 (1999).

[堀内 99] 堀内 匡, 藤野 昭典, 片井 修, 椹木 哲夫: 経験強化を考慮した *Q-Learning* の提案とその応用, 計測自動制御学会論文集, Vol.35, No.5, pp.645–653 (1999).

[Jaakkola 94] Jaakkola, T., Singh, S. P. and Jordan, M. I.: Reinforcement Learning Algorithm for Partially Observable Markov Decision Problems, *Advances in Neural Information Processing Systems 7 (NIPS-94)*, pp.345-352 (1994).

[木村 96] 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).

[木村 97] 木村 元, Kaelbling, L. P.: 部分観測マルコフ決定過程下での強化学習, 人工知能学会誌, Vol.12, No.6, pp.822–830 (1997).

[木村 99] 木村 元, 宮崎 和光, 小林 重信: 強化学習システムの設計指針, 計測と制御, Vol.38, No.10 (掲載予定).

[McCallum 95] McCallum, R. A.: *Instance-Based Utile Distinctions for Reinforcement Learning with Hidden State*, Proceedings of the 12th International Conference on Machine Learning, pp. 387-395 (1995).

[宮崎 94] 宮崎 和光, 山村 雅幸, 小林 重信. 強化学習における報酬割当の理論的考察, 人工知能学会誌, vol 9, No 4, pp.104-111 (1994).

[宮崎 95] 宮崎 和光, 山村 雅幸, 小林 重信. k-確実探査法:強化学習における環境同定のための行動選択戦略, 人工知能学会誌, vol 10, No 3, pp.124-133 (1995).

[宮崎 97a] 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.78-89 (1997).

[宮崎 97b] 宮崎 和光, 小林 重信. 離散マルコフ決定過程下での強化学習, 人工知能学会誌, Vol.12, No.6, pp.3-13 (1997).

[宮崎 98] 宮崎 和光, 小林 重信. *POMDPs* における合理的政策の逐次改善アルゴリズムの提案, 第 25 回知能システムシンポジウム予稿集, pp.87-92 (1998).

[宮崎 99a] 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).

[宮崎 99b] 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 (掲載予定).

[Samuel 59] Samuel, A. L. *Some Studies in Machine Learning Using the Game of Checkers*, IBM Journal on Research and Development 3. pp.210-229 (1959).

[Sutton 88] Sutton, R. S. *Learning to Predict by the Methods of Temporal Differences*, Machine Learning 3, pp.9-44 (1988).

[Sutton 98] Sutton, R. S. & Barto, A.: *Reinforcement Learning: An Introduction*, A Bradford Book, The MIT Press (1998).

[Singh 94] Singh, S. P., Jaakkola, T. and Jordan, M. I.: *Learning Without State-Estimation in Partially Observable Markovian Decision Processes*, Proceedings of the 11th International Conference on Machine Learning, pp. 284–292 (1994).

[山村 95] 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, pp.23-29 (1995).

[ワグナー 78] ワグナー (高橋 幸雄, 森 雅夫, 山田 尭 訳). 「オペレーションズ・リサーチ入門 5ニ確率的計画法」, 培風館, (1978).

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

[Whitehead 91] Whitehead, S. D.: *A Complexity Analysis of Cooperative Mechanisms in Reinforcement Learning*, Proc. of 9th National Conference on Artificial Intelligence, Vol. 2, pp.607-613 (1991).

著者紹介

宮崎 和光 (正会員) は, 前掲 (Vol.14, No.1, p.156) 参照.

木村 元 (正会員) は, 前掲 (Vol.14, No.1, p.130) 参照.

小林 重信 (正会員) は, 前掲 (Vol.14, No.1, p.156) 参照.