論　文

Technical Papers

POMDPs 環境下での決定的政策の学習

Learning Deterministic Policies in Partially Observable Markov Decision Processes

宮崎　和光*　荒井　幸代*　小林　重信*
Kazuteru Miyazaki　Sachiyo Arai　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.

1997年10月29日　受理

Keywords: reinforcement learning, profit sharing, Q-learning, POMDPs, rational policy making.

Summary

Partially Observable Markov Decision Process (POMDP) is a representative class of non-Markovian environments, where agents sense different environmental states as the same sensory input. We recognize that full implementation of POMDPs must overcome two deceptive problems. We call confusion of state values a *Type 1 deceptive problem* and indistinction of rational and irrational rules a *Type 2 deceptive problem*. The Type 1 problem deceives Q-learning, the most widely-used method in which state values are estimated. Though Profit Sharing that satisfies *Rationality Theorem* [Miyazaki 94] is not deceived by Type 1 problem, it cannot overcome a Type 2 problem.

A current approach to POMDPs is classified into two types. One is the *memory-based approach* that uses histories of sensor-action pairs to divide partially observable states. The other is to use *stochastic policy* where the agent selects action stochastically to escape from partially observable states. The memory-based approach needs numerous memories to store histories of sensor-action pairs. Stochastic policy may generate unnecessary actions to acquire rewards.

In this paper, we propose a new approach to POMDPs. For the subclass environment that does not need stochastic policy, we consider to learn a *deterministic* rational policy to avoid all states that manifest a Type 2 problem. We claim that the weight as an evaluation factor of a rule has the possibility to derive an irrational policy due to Type 2 problem. Therefore, no weight is used to make a rational policy. We propose the *Rational Policy Making* algorithm (RPM) that can learn a rational policy by direct acquirement of rational rules from that rule's definition. RPM is applied to maze environments. We show that RPM can learn the most stable rational policy in comparison with other methods.

1. はじめに

近年，強化学習研究のトピックスは**部分観測マルコフ決定過程** (POMDPs) に移行しつつある．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，すなわち**不完全知覚状態**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．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** [Jaakkola 94, 木村 96, Singh 94] 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で，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．

本論文では，決定的政策により合理性が保証されるPOMDPsにおける学習を考える．このクラスをTD法 [Sutton 88] やQ-learning [Watkins 92] などの状態の価値を利用するタイプの学習法が苦手とするクラス，Profit Sharing [Grefenstette 88] も含めた重みを使用した手法が苦手とするクラスの2つに分類する．重みの使用は学習の必要条件ではなく，誤った学習を引き起こす可能性があることを示唆し，重みを使わない手法の重要性を説く．

以下，**2章**では，POMDPsの困難さを2種類の**騙し**という形で分類し，問題の所在を明らかにする．**3章**では，**2章**で述べた問題点を克服し，重みを使わずに合理的政策を形成するための手法である**合理的政策形成アルゴリズム**を提案する．**4章**は数値例であり，提案手法の有効性を確認する．**5章**は結論であり，本研究の成果を総括し，今後の課題をとりまとめる．

2. 問題設定

2・1 準備

学習器は環境からの感覚入力に対し，行動を選択し，実行に移す．一連の行動に対して，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** (episode) と呼ぶ．例えば図1の環境で学習器が [xb, xa, ya, za, yb, (報酬), xa, zb, xa, yb, (報酬)] と行動したとすると，この中には (xb, xa, ya, za, yb), (xa, zb, xa, yb) の2つのエピソードが含まれている (図2参照)．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) という．例えば図2のエピソード1には (xb), (ya, za) の2つの迂回系列が含まれている．

各感覚入力に対し，選択すべき行動を1つだけ与える関数を**決定的政策** (deterministic policy)，いくつかの行動を確率的に与える関数を**確率的政策** (stochastic policy) と呼ぶ．本論文では，決定的政策のみを考える．

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** (rational policy) という．ある合理的政策の

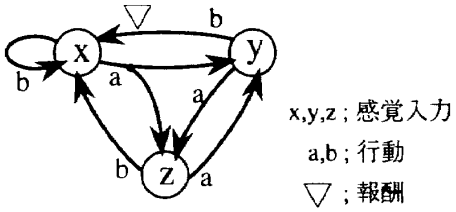

図1 例で用いた環境

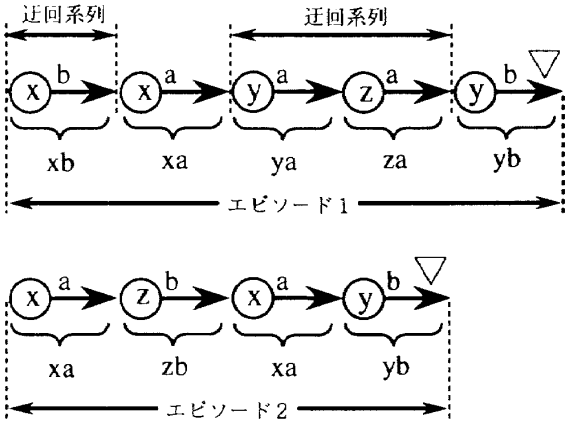

図2 エピソードおよび迂回系列の例

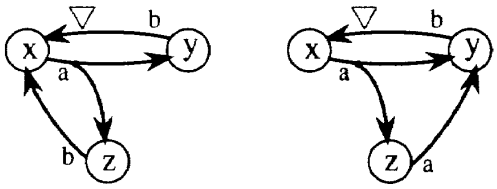

図3 図1の環境における合理的政策

構成要素となっているルールを**合理的ルール** (rational rule)，それ以外のルールを**非合理的ルール** (irrational rule) と呼ぶ．図1の環境では，図3に示すような2種類の合理的政策が考えられるので，xa,yb,za,zbの4つが合理的ルールであり，xbとyaが非合理的ルールである．

あるエピソードにおいて迂回系列上に存在しないルールは合理的である．図2ではxa,yb,zbがこれに該当する．迂回系列上に存在していても，同じエピソードにおいて同一の感覚入力に対して異なるルールが存在しなければ，そのルールは合理的である．図2ではzaがこれに該当する．

2・2 POMDPsにおける混同と騙し

POMDPsでは不完全知覚により実際には異なる環境の状態 (またはルール) が学習器にとって同一の感覚入力 (またはルール) として知覚される場合がある．このとき**混同** (confusion) が生じているという．混同により合理的政策の学習が妨げられるとき，その学習器

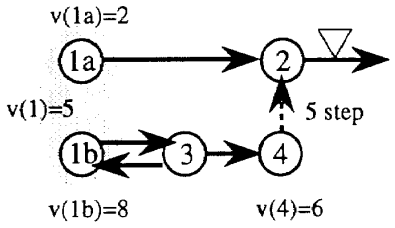

図4　タイプ 1 の混同による騙しの例

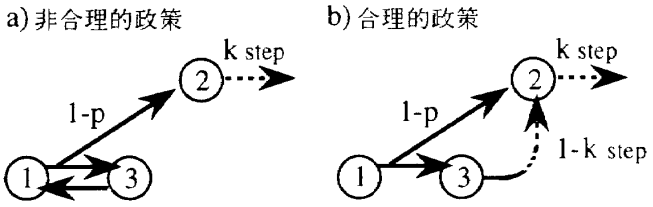

図5　図4の環境における 2 つの政策

は**騙された** (deceived) という．混同には 2 通りのタイプがあり，それぞれに騙しが起こり得る．

〔1〕 **タイプ 1 の混同と騙し**

不完全知覚により，価値の高い状態と価値の低い状態が同一視されるとき，**タイプ 1 の混同**という．

例えば，図 4の環境を考える．ここでは状態の価値を報酬への最短ステップ数で見積もるものとする．このとき状態 1a の価値は 2, 状態 1b の価値は 8 である．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る．しかし状態 1 ではつねに右へ向かう行動が最適なので，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．さらに，採用する行動選択器の種類によっては，状態 1a と 1b を等確率で経験するとは限らず，非合理的な政策と合理的な政策の間で政策が振動する可能性もある．

タイプ 1 の混同により騙しが生じる確率を図 4および図 5を利用して見積もる．今，状態 2, 3 から報酬への最短ステップ数の期待値をそれぞれ k, ℓ $(k < \ell)$ とおく．学習器が状態 1 を知覚したとき，それが状態 1b である確率を p $(0 < p < 1)$ とおく．図 4の環境では，状態 3 で左を選ぶか右を選ぶかで図 5に示すような 2 通りの政策が考えられる．政策 a) および b) における状態 1 の価値は，それぞれ $k+(1+p)/(1-p), (\ell-k)p+(k+1)$ となる．ここで政策 a) は学習されるべきでない．なぜならひとたび状態 1b に陥ると，状態 1b と 3 の間を往復し続けてしまうためである．そのような学習は，政策 b) における状態 1 の価値が政策 a) のそれよりも低いとき，すなわち以下の確率で状態 1b を経験する場合に生じる．

$$0 < p < 1 - \frac{2}{\ell - k} \qquad (1)$$

右辺は ℓ が k に比べ十分大きいときほぼ 1 となる．これは任意の p に対し非合理的な政策である政策 a) がつねに学習されることを意味する．したがってタイプ 1 の混同に起因する騙しは，状態の価値を利用する方法では避けられない問題であるといえる．

以上より，TD 法 [Sutton 88] や Q-learning (QL) [Watkins 92] など状態の価値を利用する手法では，タイプ 1 の混同による騙しにより，合理的政策が学習できない場合がある．一方，エピソード上のルール群を一括強化する Profit Sharing (PS) [Grefenstette 88] は，状態の価値を利用しないので，タイプ 1 の混同による騙しは生じないことに注意されたい．

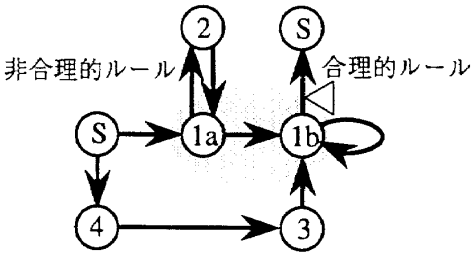

図6　タイプ 2 の混同による騙しの例

〔2〕 **タイプ 2 の混同と騙し**

不完全知覚により，合理的ルールと非合理的ルールが同一視されるとき，**タイプ 2 の混同**という．

例えば，図 6の環境を考える．この環境では，状態 1a で上という行動は非合理的ルールであるが，同じ行動は状態 1b では合理的ルールとなる．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，学習器にとっては合理的ルールとされる．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．さらに，採用する行動選択器の種類によっては，タイプ 1 の混同の場合と同様に，政策が振動する可能性もある．

タイプ 2 の混同に騙されないためには，ルールを非合理的ルールにする状態を政策から排除すればよい．しかしそのような学習は，ルールの価値を報酬への最短ステップ数で評価するような手法では，一般に困難である．今，図 6の環境では，状態 1 で上という行動を出力した場合，状態 S で右を選ぶか下を選ぶかで図 7に示すような 2 通りの政策が考えられる．このとき，状態 S で右を選ぶ行動の価値は 3 だが，下を選ぶ行動

の価値は4となる．したがって，ルールの評価に重みを利用して徐々に政策を改善していく手法では，より報酬に近い状態Sで右を選ぶ行動が強化されてしまう．

以上より，QLやPSなどルールの重みを利用してより報酬に近いルールを徐々に強化していく手法では，タイプ2の混同による騙しにより合理的政策が学習できない場合がある．

2・3 従来の接近法

従来の接近では，混同と騙しのタイプを区別せずに不完全知覚に対する対症療法的解決が試みられてきた．そのような手法は以下の2つに類別される．

〔1〕 メモリーベース法による接近

メモリーベース法は，過去の履歴を用いて，不完全知覚を生じている状態を分離する接近である．状態の分離には通常，統計的検定が利用されるため，有意な結果を得るには過去の膨大な量の履歴をメモリーに保存しておく必要がある．代表的なメモリーベース法として**Utile Suffix Memory** (USM) [McCallum 95]がある．

USMでは，過去の履歴を木構造で表現し，各葉ノードを内部状態とすることにより，可変長の履歴を扱う．ここで参照すべき過去の履歴の長さは，**fringe**というパラメータで制御される．fringeが短すぎると，騙しを生じている状態を正しく分離しきれない場合がある．そのためfringeは十分長くとる必要があるが，USMが要する記憶容量は，学習器が観測する状態の種類をnとすれば，行動の種類を無視したとしても最悪の場合，$O(n^{fringe})$となり，一般に膨大な量のメモリーを必要とする．

〔2〕 確率的政策法による接近

確率的政策では，確率的に行動を選択することにより，騙し状態からの脱出を試みる．[Jaakkola 94, 木村 96]らは，確率的政策を学習する強化学習手法を提案している．

[木村 96]の**確率的傾斜法**では，今までとった行動の情報論的な意味での価値を，**適正度の履歴** (eligibility trace)という形で，過去の行動ほど割り引いて記憶していく．報酬を得た時点で，この適正度の履歴を利用しルールの重みに相当する内部変数を更新する．その結果，割引報酬の期待値を最大化する方向へと政策を確率的に逐次改善することが可能となる．

確率的政策では，行動選択に確率が含まれるため，報酬を得る際に要する行動数は，乱数の振られ具合に大きく依存してしまう．さらに，確率的政策を必要としない状態に対しても確率的に行動を出力してしまう欠点を持つ．

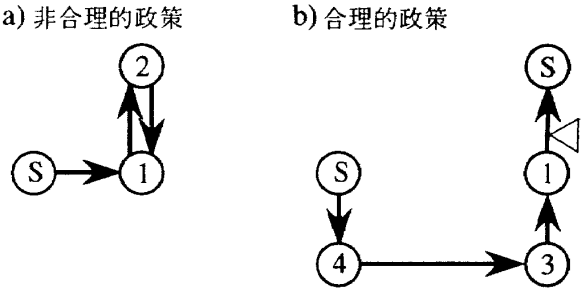

図7 図6の環境における2つの政策

2・4 本論文の立場および接近法

従来の接近法は，タイプ1またはタイプ2の混同による騙しに直面したとき，そのような状態を分離すること(メモリーベース法)またはそのような状態から確率的に脱出すること(確率的政策法)を試みている．本論文では，混同による騙しに陥らないように，事前にそれを回避することを考える．

本論文では，混同による騙しを回避する決定的政策が存在するPOMDPsのサブクラスを対象として，合理的かつ決定的な政策を効率良く形成する方法を提案する．

まず，タイプ1の混同による騙しを回避するために，状態の価値は利用しないものとする．タイプ2の混同による騙しの下では，1つのルールが合理的ルールであったり，非合理的ルールであったりする．本論文では，ルールを非合理的ルールにする状態を政策から排除することにより，タイプ2の混同による騙しを回避することを考える．例えば，図6の騙しの例に対しては，図7b)に示される合理的政策を学習することを考える．

2・2節で議論したように，Q-learningのように状態の価値を利用する方法ではタイプ1およびタイプ2の混同による騙しを回避することは困難である．また，Profit Sharingのように，状態の価値は利用しないが，ルールの重みを利用する方法ではタイプ2の混同による騙しを回避することは困難である．

そこで，本論文では，状態の価値およびルールの重みを使わずに，合理的ルールから，直接，合理的政策を形成することを考える．**次章**では，合理的政策を形成するための具体的アルゴリズムを提案する．

3. 合理的政策形成アルゴリズム

3・1 基 本 定 理

〔1〕 合理的ルールの抽出

あるエピソードにおいて同一の感覚入力に対して，報

酬に最も近い位置で選択されていたルールは，定義より合理的ルールである．例えば図2のエピソード1からは，ルールyb, za, xaが合理的ルールと判断される．

また逆に，あるルールxaが合理的ルールであるならば，感覚入力xの中でも，最も報酬に近い位置で行動aが選択されているようなエピソードが存在するはずである．例えば図1の環境では，ルールzaは合理的ルールであるが，行動aが最も報酬に近いzで選択されているエピソードとしてエピソード1が存在する．

以上より，あるルールxaが合理的ルールであるための必要十分条件は以下のように与えられる．

［定理 1］（合理的ルールの抽出）
ルールxaが合理的ルールであるための必要十分条件は，あるエピソードにおいて同一の感覚入力xに対する行動選択の中でaが報酬に最も近い位置に存在することである． □

定理1は，タイプ1およびタイプ2の混同と騙しが存在していても成立する．定理1は状態の価値を利用しないので，タイプ1の混同と騙しの影響を受けない．タイプ2の混同が生じているルールは定理1により合理的ルールとして抽出されることに留意されたい．

〔2〕 合理的政策の形成

まず，タイプ2の混同が生じない場合を考える．定理1を繰り返し適用して合理的政策を形成することを考える．最初に，任意のエピソードが1つ与えられたものとする．定理1を適用することにより，合理的ルールを抽出する．次に，抽出された合理的ルールを優先して使うことにより，次のエピソードをつくる．このエピソードに対し定理1を適用して合理的ルールを抽出する．以下同様のことを繰り返せば，各エピソードごとに報酬が得られるので，結果として合理的政策が得られる．

また逆に，定理1によって抽出された合理的ルールを優先して選択しない場合，合理的政策が形成されないことがある．例えば，図8の環境を考える．この環境では，すべてのルールは合理的であるが，ルールxaとyaが，同一のエピソードから抽出されることはない．定理1により抽出された合理的ルールを優先して選択した場合，この2つのルールが同時に選択されることはないが，定理1による抽出を優先せず，あえてこれらのルールを選択し続けた場合，報酬は一切得られない．よって，つねに合理的政策が形成されるためには，定理1により抽出された合理的ルールを優先して選択することが必要である．

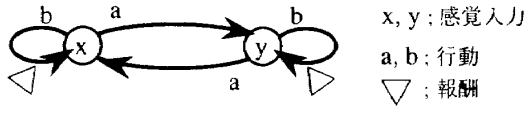

図8 合理的ルールのみで構成される環境の例

以上の議論より，次の定理が得られる．

［定理 2］（合理的政策の形成）
タイプ2の混同が存在しないクラスにおいて合理的政策がつねに形成されるための必要十分条件は，定理1によって抽出された合理的ルールを優先して選択することである． □

タイプ2の混同が存在するクラスでは，定理2を使って合理的なルールを優先して選択したとしても，騙しにより必ずしも合理的政策が形成されるとは限らない．合理的政策の形成に失敗したと判断される場合には，新しいエピソードを任意に生成して政策の形成をやり直すマルチスタート法を採用する必要がある．

3・2 合理的政策形成アルゴリズムの提案

定理1と定理2によってつくられる**合理的政策形成**アルゴリズムを以下に示す．

```
procedure 合理的政策形成(Rational Policy Making)
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.
```

まず学習器には予め1次と2次合わせて合計$2N$個分の記憶領域を確保しておく．ここでNは感覚入力の種類である．Nの値は学習以前には既知でなくてもよい．学習器は行動を出力するごとに，1次記憶上に，そのとき出力した行動を上書きする．以上を報酬が得られるまで繰り返し，報酬が得られた時点で，1次記憶領域の内容を2次記憶領域に複写する．

この結果，2次記憶領域には，定理1に従い合理的ルールのみが記録されていく．学習器は，合理的ルールが判明している感覚入力を知覚した場合には，その

ルールに従って行動を出力し，そうでない場合には，環境を探査するための行動を出力する．

環境探査戦略として，ランダム探査，k-確実探査法[宮崎 95] などが考えられる．環境が MDPs であれば，k-確実探査法による効率的な探査が期待できるが，POMDPs 環境ではランダム探査が推奨される．

2 次記憶には政策が保存されるので，その収束を判定すれば，学習の打ち切りが可能となる．収束の判定としては，2 次記憶が最後に更新されたステップ数を n とし，その 2 倍の $2n$ ステップ行動を出力しても 2 次記憶が更新されなければ，収束したとみなす方法が例えば考えられる．なお，状態遷移が決定的な MDPs 環境の場合，報酬を得た後に状態遷移した先の 2 次記憶が埋まっていれば，合理的政策が形成されたと判断してよい．

タイプ 2 の混同が存在しないクラスでは，上記の収束判定後，必ず合理的政策が得られる．しかし，タイプ 2 の混同が存在する場合には，得られた政策が合理的政策になっているとは限らない．収束判定期間内に報酬が全く得られなければ，合理的政策の学習に失敗したと判断される．この場合，1 次および 2 次記憶の内容を初期化して，新しい政策の学習を行うものとする．

3・3 動　作　例

図 9の環境を用いて動作例を示す．まず最初は 1 次記憶，2 次記憶ともに初期化する．最初，2 次記憶は空なので，ランダムに xb, ya, zb, wb という順番でルールが選ばれたとする．その結果，対応する 1 次記憶の感覚入力上に行動が記述される．次にルール wa が選択された際には，対応する 1 次記憶，すなわち w 上に行動 a が上書きされる．同様にして ya, za という順番でルールが選ばれ，1 次記憶が更新される．

ルール za を実行した後，報酬が得られた．すると 1 次記憶の内容が 2 次記憶にコピーされる．すると次からは 2 次記憶に行動が記述されている感覚入力を知覚した場合には 2 次記憶上の記述に基づき行動が出力される．今，この例では，感覚入力 x を知覚しているので，必ず行動 b が出力される．同様に ya, za が選ばれ，報酬が得られた．このように既に知覚した感覚入力に関しては，たちどころに合理的ルールが判明し，継続的な報酬の確保が実現される．

さて，同様に，xb が選ばれた後，今まで知覚したことのない v という感覚入力を知覚した．この場合，ランダムに行動 a が選ばれ，その行動が 1 次記憶上に記述される．その後 y に遷移したため，以後は 2 次記憶の内容に基づき ya, za というルールが選ばれ，報酬が得られた．このとき今まで空白だった 2 次記憶上の v の領域が埋まる．その結果，環境中のすべての状態遷移を経験したことになり，合理的政策が形成されたこととなる．さらに最適性を追求するために，1 次記憶および 2 次記憶を初期化し，同様の処理を繰り返す．

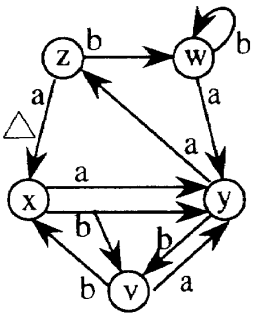

図 9　動作例で用いた環境

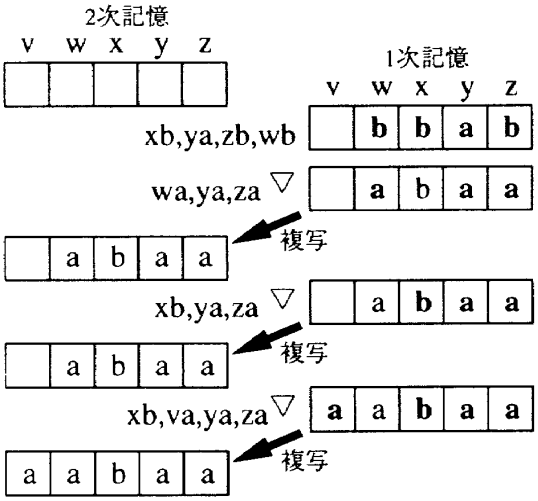

図 10　合理的政策形成アルゴリズムの動作例

3・4 合理的政策形成アルゴリズムの特徴

合理的政策形成アルゴリズムは，決定的な合理的政策が存在する環境下では，それらのうちのいずれか 1 つを必ず学習することができる．

さらにタイプ 2 の混同が存在しないクラスでは，1 回の試行で必ず合理的政策が得られる．ここで試行とは，1 次および 2 次記憶の初期化から政策の収束判定終了までをいう．

また，一般に，1 回の試行に要する行動数は非常に少ない．そのため，1 次および 2 次記憶を初期化することで，より優れた政策を発見することが可能となる．

一方，タイプ 2 の混同が存在するクラスでは，試行を繰り返すことで，騙しを回避する政策を発見する．1 試行に要する行動数が少ないので，このようなマルチスタート法が有効である．

4. 迷路走行問題による性能評価

4・1 Q-learning および Profit Sharing との比較

合理的政策形成アルゴリズム (RPM) を図 11に示す迷路的環境を用いて Q-learning (QL) および Profit Sharing (PS) と比較する．学習器は各マス目で上，下，左，右の中から 1 つの移動を選択できる．始点は S であり，終点 G に到着すれば報酬が与えられる．視界は学習器の周囲 8 マスに限定される．例えば状態 1a と 1b は学習器にとっては同一の状態 (状態 1) として知覚される．図 11a) ではタイプ 1 の混同もタイプ 2 の混同も生じていないが，図 11b) では，例えば，状態 2 でタイプ 1 の混同が生じ，図 11c) では状態 1 でタイプ 2 の混同が生じている．

図 11の各迷路に対し，RPM を他手法と比較した結果を表 1～表 3にまとめる．RPM では合計 100 個の合理的政策を生成し，他手法では乱数の種を変えて行った 100 回の実験の結果を示している．表には，そのうちの合理的政策が獲得された回数をその内訳とともに示している．ここで例えば 11 ステップとは，S から G への行動数が 11 である政策を意味する．QL の学習率は 0.05，割引率は 0.95 である．これらの値は予備的実験の後決定した．PS では，[宮崎 94] の定理を満たす公比 0.2 の等比減少関数を用いルールを強化した．QL および PS の初期重みは 10.0，報酬値は 100.0 である．RPM では，合理的政策が得られなかった場合，その政策を破棄し再試行する．政策の破棄は図 11c) においてのみ生じ，その回数は 314 回であった．

図 11a) にはタイプ 1 およびタイプ 2 の混同が存在しないため，QL ではつねに最適である 11 ステップの政策が学習された．PS は本来最適性は保証されないが，重みによる政策の改善が有効に機能し，100 個の政策すべてが最適政策となった．RPM はそもそも最適性を追求する手法ではないが，最適政策が最も多く学習された．

図 11 b) では，QL は状態 2 におけるタイプ 1 の混同につねに騙された．すなわち状態 2a と 2b を同じ状態 (状態 2) と知覚したことで，状態 2 の価値が状態 15 の価値よりも高くなり，状態 2a と 11 の間を往復する非合理的な政策が学習された．PS および RPM は状態の価値を利用しないので，このようなタイプ 1 の混同に騙されず，つねに合理的政策が学習された．また PS ではつねに最適である 8 ステップの政策が学習された．これは図 11a) の場合と同様に，重みを使用した解の逐次更新が有効に機能した結果である．

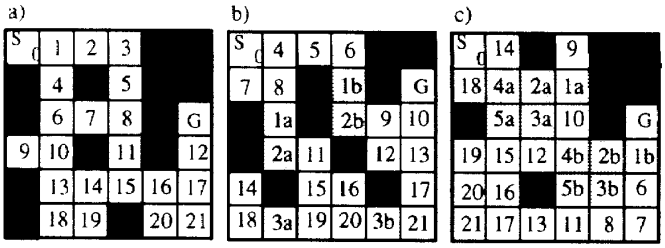

図 11 性能評価で用いた迷路環境．
a) タイプ 1 およびタイプ 2 の混同が存在しない環境，
b) タイプ 1 の混同が存在する環境，
c) タイプ 2 の混同が存在する環境

表 1 図 11a) における合理的政策の獲得回数

	合理的政策	合理的政策の内訳				
		11step	13step	15step	17step	19step
QL	100	100	0	0	0	0
PS	100	100	0	0	0	0
RPM	100	70	19	7	3	1

表 2 図 11b) における合理的政策の獲得回数

	合理的政策	合理的政策の内訳				
		8step	10step	12step	14step	16step
QL	0	0	0	0	0	0
PS	100	100	0	0	0	0
RPM	100	36	22	3	28	11

表 3 図 11c) における合理的政策の獲得回数

	合理的政策	合理的政策の内訳			
		11step	13step	15step	17step
QL	0	0	0	0	0
PS	0	0	0	0	0
RPM	100	56	32	10	2

タイプ 2 の混同が存在する図 11c) では，QL のみならず PS までも学習できていない．これは重み使用の弊害のためである．一般に重みを使用すると，報酬から遠い状態の学習が遅れる．図 11c) では，状態 1a を回避する政策が要求されるが，最短でゴールに向かうためには，この状態を通過すべきである．このような場合，重みを使用した手法では，状態 1a に接近する学習が徐々になされてしまい，結果的に，状態 1a と 9 の間を往復する非合理的な政策が学習されてしまう．一方，RPM では，状態 1a を回避する政策をひとたび発見すれば，それを維持し続ける．そのため重みの弊害はなく，つねに合理的政策を保持することが可能となる．

4・2 POMDPs を対象とする手法との比較

次に，POMDPs を対象とする手法である Utile Suffix Memory (USM) [MaCallum 95] および確率的傾

表4 図11a)における解の質および学習に要した行動数

	解の質		学習に要した行動数	
	平均	標準偏差	平均	標準偏差
RPM	11.9	0.87	4.72×10^2	4.27×10^2
SGA	11.0	0.00	2.11×10^4	4.44×10^3

表5 図11b)における解の質および学習に要した行動数

	解の質		学習に要した行動数	
	平均	標準偏差	平均	標準偏差
RPM	11.1	2.96	3.01×10^2	2.97×10^2
SGA	8.0	0.00	1.15×10^4	3.20×10^3

表6 図11c)における解の質および学習に要した行動数

	解の質		学習に要した行動数	
	平均	標準偏差	平均	標準偏差
RPM	12.1	1.50	1.83×10^3	1.92×10^3
SGA	12.4	0.91	2.04×10^4	5.47×10^3

斜法 (SGA)[木村 96] との比較を行う.

USMは，例えば図11c)の場合，状態1aと1bを区別するために4ステップ前に位置していた状態の情報を必要とする．したがってそこでは，行動を木構造に含まなかったとしても，最低fringe=4，すなわち22^4もの状態表現を必要とする．USMに代表されるメモリーベース法は，一般に，このように膨大な量のメモリーを必要とする．そのため現実的な手法とは言い難い．

図11の各迷路におけるRPMとSGAの比較結果を表4～表6にまとめる．ここで解の質とは，学習の結果得られた政策によるSからGまでのステップ数である．RPMでは合計100個の合理的政策を生成し，SGAでは乱数の種を変えた100回の実験の結果を示している．SGAの学習率は0.1，割引率は0.95である．これらの値は予備的実験の後決定した．

SGAは図11a)およびb)ではつねに最適政策が学習された．しかしその学習に要する行動数はRPMの約100倍である．RPMでは，政策を次々生成し，その中で最良のものを選択する形で最適性を保証することができる．表1および表2より，RPMでは100回中30～40回以上，最適政策が得られている．したがってRPMは，SGAの約1/30の行動数で最適政策を学習できたことになる．

図11 c)では，学習に要した行動数のみならず解の質に関しても若干ではあるがRPMがSGAを上回った．SGAは本迷路においては，100回中，11ステップの決定的政策が27回，状態1で上と下を確率的に出力する確率的政策が67回，状態2で右と下を確率的に出力するなどの確率的政策が6回獲得された．状態1で上と下を確率的に出力した場合，最少ケースでは9ステップで報酬を得ることができるが，平均では13ステップとなる．SGAでは，確率的政策の9ステップという最少ケースに惑わされ，11ステップである最適政策が多くの場合破壊されている．確率的政策は決定的政策で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，このように不用意に確率的政策が学習されてしまう場合がある．ここに確率的政策の限界がある．

5. おわりに

本論文では，決定的政策で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**を提案した．そこでは，対象問題クラスを，TD法やQ-learningなどの状態の価値を推定するタイプの学習法が苦手とするクラス，およびProfit Sharingも含めた重みを使用した手法が苦手とするクラスの2つに分類し，問題の所在を明らかにした．

従来，POMDPsに対する接近としては，膨大な量のメモリーを要するメモリーベース法や確率的政策を用いる方法などが存在した．提案手法は，非常に少ない量のメモリーで学習でき，また，つねに特定の行動が出力されるので確率的政策でみられるような不安定な挙動を示すこともない．迷路環境に適用することで，提案手法は，従来のいずれの手法よりも安定して合理的政策を学習できる手法であることを示した．

本論文ではPOMDPsを対象としたが，環境がMDPsであれば，ここで提案した合理的政策アルゴリズムに，k-確実探査法やPolicy Iteration Algorithm (PIA) [ワグナー 78] を組み合わせることで，より直接的に最適政策を形成することができる．

今後の課題としては，1) 政策の逐次改善法の構築，2) 多種類報酬への拡張，3) マルチエージェント系への適用，などが挙げられる．

◇参考文献◇

[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).

[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).

[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).

[Singh 94] Singh, S. P.: Reinforcement Learning Algorithms for Average-Payoff Markovian Decision Processes, *Proceedings of the 12th National Conference on Artificial Intelligence*, Vol. 1, pp. 700–705 (1994).

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

[ワグナー 78] ワグナー（高橋幸雄，森 雅夫，山田 竟 訳）：「オペレーションズ・リサーチ入門 5=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).

〔担当委員：安倍直樹〕

著 者 紹 介

宮崎 和光（正会員）

1991 年明治大学工学部精密工学科卒業．1996 年東京工業大学大学院総合理工学研究科知能科学専攻博士後期課程修了．工学博士．同年 4 月，同大学大学院総合理工学研究科助手．1998 年 4 月，同大学大学院総合理工学研究科リサーチアソシエイト．現在に至る．人工知能，特に強化学習に関する研究に従事．計測自動制御学会，日本機械学会各会員．〈teru@fe.dis.titech.ac.jp〉

荒井 幸代（正会員）

1984 年慶応義塾大学工学部計測工学科卒業．(株)ソニー，筑波大学大学院準研究員，米カリフォルニア大学バークレー校客員研究員を経て，1998 年東京工業大学大学院総合理工学研究科知能科学専攻博士後期課程修了．工学博士．現在，同大学大学院総合理工学研究科リサーチアソシエイト．マルチエージェントの協調に関する研究に従事．AAAI 学会会員．〈arai@fe.dis.titech.ac.jp〉

小林 重信（正会員）は，前掲（Vol.14, No.1, p.130）参照．