社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－MPS－48　(9)
2004／3／2

確率的な遷移を含んだ部分観測マルコフ決定過程における強化学習法

長田 浩　藤田 聡

Wiering らによって提案された HQ 学習は，タスクを複数のマルコフ的なサブタスクに分割し，それぞれを独立に学習することで部分的な観測によって起こる知覚の見せかけ問題を解決している．しかしマルチエージェント強化学習において状態遷移は一般に確率的に起こるが，HQ 学習ではその枠組みのために適切な学習がされない場合がある．本稿ではこの問題を解決するために HQ 学習を拡張した手法を提案し，その性能を実験的に評価する．

Reinforcement Learning in Partially Observable Markov Decision Process Including Probability State Transitions

Hiroshi OSADA　Satoshi FUJITA

HQ-learning proposed by Wiering *et al* decomposes a given task into several independent Markovian subtasks, and activates those tasks in a sequential manner. However, in multi-agent systems which have probability state transitions, HQ-learning cannot learn appropriately due to the architecture. In this paper, we propose a new learning scheme to solve such problems, and evaluate the effectiveness experimentally.

1　はじめに

近年，複数の自律的なエージェントが協調することでより複雑な問題の解決を目指すマルチエージェントシステム (MAS) が注目されている．しかし，MAS において，その問題に対応した政策を人間が完全に記述することは困難である．そこで試行錯誤を通じて環境に適応する機械制御システムである強化学習を用いて政策の獲得を行う研究が盛んに行われている．強化学習とは，移動等の行為を行うエージェントが，直接の教師を持たずに，行為に対する環境からの報酬と呼ばれるスカラー情報だけから学習を行う自律的学習である．このため，多くの問題に適用可能であり，また，試行錯誤を通じた学習であるので，人間の経験則に反する優れた政策を発見することが期待される．

強化学習ではこれまでマルコフ決定過程 (Markov Decision Process: MDP) としてモデル化できる環境を対象とする研究が広く行われてきた [1]．しかし，実問題では，センサの能力が不十分である場合や，完全な知覚を得ることが可能であっても状態数の爆発を抑えるために情報を制限する場合が考えられる．このため，MDP の状態観測に不完全性を付加した部分観測マルコフ決定過程 (partially observable MDP: POMDP) としてモデル化できる環境を対象とした学習方法が研究されている．Wiering らはこの不完全性を克服するために，タスクを複数のサブタスクに分割し，そのサブタスクを順序付けされたサブエージェントにそれぞれ独立に学習させる HQ 学習を提案した [3]．HQ 学習は特定のクラスの POMDP のタスクを効率的に学習できるが，そのタスクが確率的な状態遷移を含む場合，適用することができないこ

広島大学大学院工学研究科情報工学専攻
〒 739-8527 東広島市鏡山 1-4-1
Graduate School of Engineering, Hiroshima University
Kagamiyama 1-4-1, Higashi-Hiroshima, 739-8527 Japan

とがある．MAS においては各エージェントが独立に学習し，それらの行動が他のエージェントの観測に影響するため，あるエージェントのある行動による状態遷移は確率的に起こると考えるべきである．

そこで本稿では，HQ 学習において固定されていたサブエージェントの順を任意の順にすることで，確率的な状態遷移を含んだ POMDP を扱う学習法を提案する．提案手法は HQ 学習に基づいており，HQ 学習における HQ テーブルを拡張することでこれを実現する．この拡張により，HQ 学習では適切な政策を表現できないタスクでも扱えるようになることが確認できた．また，マルチエージェント系に適用した場合の性能についても実験的に評価する．

2 POMDP

本稿で対象とする POMDP を $\langle S, s_1, A, P, R, O, B, \gamma \rangle$ の組として表わす．ここで S は有限の状態集合，$s_1 (\in S)$ はエージェントの初期集合，A はエージェントの行動集合，P はエージェントの行動による状態遷移の確率を表わす関数とする．すなわち，状態 $s \in S$ においてエージェントが行動 $a \in A$ を実行し，状態が確率的に $s' \in S$ に遷移する場合の遷移確率は $Pr\{s_{t+1} = s' | s_t = s, a_t = a\} = P_a(s, s')$ により表わされる．状態遷移の確率はその状態以前の遷移の系列には依存しない．R は状態遷移に対してエージェントに与えられる報酬の期待値とする．先の P の説明に加え，このときに環境からエージェントに報酬 r_t が与えられた場合の期待値は $E\{r_t | s_t = s, a_t = a, s_{t+1} = s'\} = R_a(s, s')$ により表わされる．O は有限の観測集合，$B : S \to O$ は状態から観測への決定的な写像とする．γ は割引率と呼ばれ，即時報酬と将来の報酬のトレードオフ比を表わす．

一般に $|S| > |O|$ なので，$B(s_i) = B(s_j) = o(\in O)$, $s_i \neq s_j$ となる状態 s_i, s_j が存在する．エージェントは o を得ただけではそれを分類することができなく，これを**知覚の見せかけ問題 (perceptual aliasing)** [4] と言う．さらにこの 2 つの状態で選択すべき行動が異なるとより困難な問題となる．

3 HQ 学習

HQ 学習は Q 学習を階層的に拡張した学習法である．HQ 学習では，1 つのエージェントが順序付けられた複数のサブエージェントにより構成され，それぞれのサブエージェントはタスクにおける MDP のサブタスクを見つけこれを解くことを学習する．各サブエージェントは (1) 適当なサブゴール，(2) 特定のサブゴールを与えられた MDP の政策を学習する．

HQ 学習の構成を図 1 に示す．エージェントは複数のサブエージェントにより構成され，サブエージェント C_1 から C_M まで，それぞれのサブゴールに到達することで，順に制御が移される．M は予め決められたサブエージェントの数である．

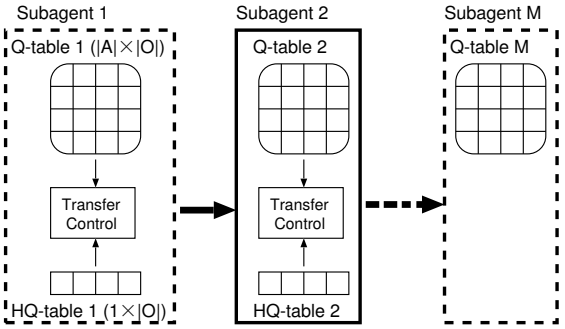

図 1: HQ 学習におけるサブエージェントの構成.

C_1 から開始し，以下の手順によって表わされる試行を繰り返すことで学習を行う．1 回の試行は離散時間ステップ $t = 1, \ldots, T(\leq T_{\max})$ により構成される．ここで T はタスク全体のゴールに到達した時間であり，$T_{\max}$ は予め設定された時間の制限である．時間 $t = T_{\max}$ になってもゴールに到達しない場合はそこでその試行を終了し，$T = T_{\max}$ となる．

1. 制御が移されたサブエージェント C_i の HQ テーブル HQ_i に基づき，Max-Uniform ルールによりサブゴール $\hat{o}_i$ を決定する．Max-Uniform ルールは確率 $Pr_{\max}$ で最大の HQ 値を持つ観測を選択し，確率 $1 - Pr_{\max}$ で全ての観測を等確率に選択する．
2. 手順 1 で選択されたサブゴール $\hat{o}_i$ に到達するまで，Q テーブル Q_i に基づき，Max-Boltzmann ルールにより選択された行動 a を実行する．
   Max-Boltzmann ルールは確率 $Pr_{\max}$ で得られた観測 o に対して最大の Q 値を持つ行動を選択し，確率 $1 - Pr_{\max}$ で以下の式 (1) で表わされるボルツマン分布に基づく確率で行動を選択する：
   $$prob_o^i(a) = \frac{e^{Q_i(o,a)/\tau}}{\sum_{a' \in A} e^{Q_i(o,a')/\tau}} \qquad (1)$$
   ただし，τ はランダム性を調整する温度パラメータである．
3. サブゴールに到達すると，$t_{i+1} = t + 1$ として次のサブエージェント $i + 1$ に制御を移し，手順 1

に戻る．ここで t_i は C_i に制御が移された時間を表わす．

エージェントがゴールに到達するか，時間が $T_{\max}$ を経過すると Q 値はオフライン Q(λ) 学習 [5] を用いて更新され，HQ 値は，その試行において最後に制御が移されたサブエージェントを C_N とすると，$C_N, C_{N-1}, \ldots, C_1$ の順で以下の規則によって更新される．

$$R_i = \sum_{t=t_i}^{t_{i+1}-1} \gamma^{t-t_i} R_a(s_t, s_{t+1}) \quad (2)$$

$$HQ'_i(\hat{o}_i) \leftarrow R_i + \gamma^{t_{i+1}-t_i}\{(1-\lambda)\max_{o'\in O} HQ_{i+1}(o') + \lambda HQ'_{i+1}(\hat{o}_{i+1})\} \quad (3)$$

$$HQ_i(\hat{o}_i) \leftarrow (1-\alpha^{HQ})HQ_i(\hat{o}_i) + \alpha^{HQ} HQ'_i(\hat{o}_i) \quad (4)$$

ここで $HQ_i(o)$ は C_i の観測 o に対する HQ 値，$\hat{o}_i$ は C_i により選択されたサブゴール，α^{HQ} $(0 < \alpha^{HQ} \leq 1)$ は HQ 値の学習率，$HQ'_i(o)$ は適合度トレースを用いた場合の望まれる値であり，λ $(0 \leq \lambda \leq 1)$ は適合度トレースを用いる程度を表す定数である．

4 提案手法

4.1 構成

提案手法では HQ 学習で $|O| \times 1$ であった HQ テーブルを $|O| \times M$ に拡張する．すなわち，サブエージェント切り替えのトリガとなる観測だけでなく，制御を移すサブエージェントも合わせて学習させることを考える．

C_1 から開始し，以下の手順により表される試行を繰り返すことで学習を行う．なお，サブエージェント C_i が観測 o_j においてサブエージェント C_k に制御を移す HQ 値を $HQ_i(o_j, C_k)$ により表わし，また，HQ 学習と同様に，タスクのゴールに到達するか時間が予め定めた $T_{\max}$ を経過することを終了条件とする．

1. 制御が移されたサブエージェント C_i は，$\forall o \in O$ に対して制御を移すサブエージェントを Max-Uniform ルールにより決定する．ただし，観測 o_j において制御を移さない場合は自分自身に制御を移す (C_i を選択する) ことでこれを表わす．また，制御が移された時点における観測 o_{t_i} に対しては，C_i を選択し，行動を実行せずに他のサブエージェントに制御を移すことを避ける．
2. 手順 1 で定めたサブゴールの組 (o_j, C_k) において，$C_k \neq C_i$ となる観測を得るまで Max-Boltzmann ルールによって選択された行動を実行する．
3. $C_k \neq C_i$ となる観測を得ると，C_k に制御を移して手順 1 に戻る．

終了条件を満たすと，Q 値は HQ 学習と同様の学習規則に従って更新する．HQ 値は，その試行において制御が移された逆順で，実行されたサブゴールについてのみ更新する．ただし，式 3 の項 $\max_{o'\in O} HQ_{i+1}(\hat{o}')$ における O をそのサブエージェントに制御が移されていた間に得た観測の集合に変更する．

4.2 HQ 学習との相違点

本研究は MAS のように確率的な状態遷移を含むタスクを対象としている．HQ 学習ではサブエージェントの順が静的であり，また唯一のサブエージェントにしか制御を移すことができないため，確率的な遷移を含む POMDP に対して適切な学習がされない場合がある．例として，HQ 学習の枠組みでは適切な決定的政策を表現することができず，HQ 学習は図 2 に示すタスクを学習できない．

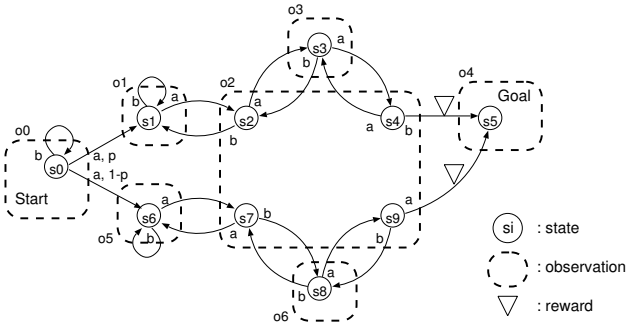

図 2: 確率的な遷移を含む POMDP の例．

提案手法ではこのタスクを学習することができる．実際に，以下のパラメータにより 1000 回の試行を 1000 回行った結果，98.5% は必ずゴールに到達する政策を得て，89.8% はステップ数 5 の最適な政策を獲得した：$T_{\max} = 1000, M = 3, \gamma = 0.9, \alpha^Q = \alpha^{HQ} = 0.1, \tau = 0.1, \lambda = 0.9$，$Pr_{\max}$ を最初の試行で 0.9 とし 1.0 まで線形に増加させる．

5 実験

5.1 タスク概要

追跡問題は MAS における強化学習の評価によく用いられている．本稿では図 3 に示す追跡問題を扱う．この実験により，マルチエージェント系における提案手法の性能を評価する．

9×9 のトーラス空間において，図 3 の状態から開始し，逃亡者が逃亡者の四方を取り囲むことを目標とする．追跡者 → 逃亡者の順で上・下・左・右・停止の 5 種類の行動のいずれかを選択する．ただし，同一のマスに複数のエージェントが重なることはできないとする．実験中は追跡者のみ学習を行い，逃亡者は学習を行わない．

各追跡者 (エージェント) が得られる観測は他の追跡者・逃亡者が (1) 自分の周囲 8 マスにいるか，(2) その周囲 16 マスにいるか，(3) 更にその周囲 24 マスにいるか，(4) それ以外にいるかを，(逃亡者，追跡者の集合) の組により表わされる．例として，図 4 の場合，中心の追跡者は $(2, \{1, 3, 4\})$ を観測として得る．実験の簡単のために $|O| = 4^4 = 256$ とした．逃亡者は空間全体を観測し，5 種類の行動後に最も近い追跡者との距離を最大化する行動を選択する．複数の行動でこの距離が等しくなる場合は，上・下・左・右・停止の順に決定的に選択する．

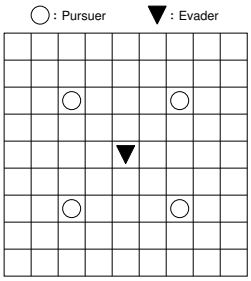

図 3: 9×9 追跡問題.

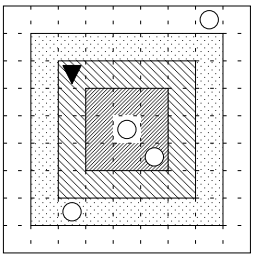

図 4: 追跡者の観測.

5.2 パラメータ設定

予備実験により，各パラメータを以下のように定める．ゴール到達時に各エージェントに 500 の報酬を与え，それ以外では -0.1 の報酬を与える．1 回の実験を 20000 回の試行により行い，$T_{\max} = 1000$ とする．提案手法に関しては，$\gamma = 0.9$，$\alpha^{HQ} = 0.001$，$Pr_{\max}$ は最初の試行では 0.8 とし，1.0 まで線形に増加させる．HQ 学習に関しては，$\gamma = 1.0$，$\alpha^{HQ} = 0.01$，$Pr_{\max}$ は 0.6 から 1.0 まで線形に増加させる．残りのパラメータは両手法とも共通に，$M = 4, \alpha^{Q} = 0.05$，$\tau = 0.2$，$\lambda = 0.9$，HQ 値，Q 値の初期値は 0.0 とする．

5.3 結果と考察

それぞれ 100 回の実験を行った．成功確率とその平均ステップ数を表 1 に示す．

提案手法は確実に目標を達成しており，この結果から，提案手法は HQ 学習に対してマルチエージェント強化学習を効率的に行えると言える．これに対して HQ 学習では過半数を超えてはいるが，提案手法と比較すると成功確率は低い．これは探査戦略をとることで他のエージェントとの同期がとれなくなった場合，これを修正できなかったためと考えられる．

表 1: 学習後の政策の性能.

	提案手法	HQ 学習
成功確率	100%	78%
平均ステップ数	3.43	4.69

6 むすび

本稿では，POMDP における MAS に対する強化学習法として確率的な遷移を考慮した強化学習法を提案した．提案手法は HQ 学習の単純な拡張であるが，HQ 学習では学習不可能な POMDP のタスクに適用できる．また実験により，POMDP としてモデル化できるマルチエージェント系に適用できることを示した．提案手法が適用可能な POMDP のクラスの明確化を今後の課題とする．

参考文献

[1] R. S. Sutton, and A. G. Barto. Reinforcement Learning: An Introduction. *The MIT Press*, Cambrigde, 1998.

[2] C. J. C. H. Watkins, and P. Dayan. Technical notes: Q-learning. *Machine Learning*, 8:279–292, 1992.

[3] M. Wiering, and J. Schmidhuber. HQ-learning. *Adaptive Behavior*, 6(2):219–292, 1997.

[4] S. D. Whitehead, and D. H. Ballard. Learning to perceive and act by trial and error. *Machine Learning*, 7:45–83, 1991.

[5] Lin. L. Reinforcement Learning for Robots Using Neural Networks. PhD thesis, Carnegie Mellon University, Pittsburgh, 1993.

[6] 山城啓秀, 上野敦志, 武田英明. 遅れ報酬に基づく遺伝的アルゴリズムによる部分観測マルコフ決定問題の解決手法. **電子情報通信学会**, J84-D-I(12):1635–1647, 2001.