時変パラメータを持つ進化的強化学習システム

非会員 梅迫 公輔* 正員 大林 正直**
 非会員 小林 邦和**

Evolutionary Reinforcement Learning System with Time-Varying Parameters

Kosuke Umesako*, Non-member, Masanao Obayashi**, Member, Kunikazu Kobayashi**, Non-member

In this paper, an evolutionary reinforcement learning system with time-varying parameters that can learn appropriate policy in dynamical POMDPs is proposed. The proposed system has time-varying parameters that can be adjusted by using reinforcement learning. Hence, the system can adapt to the time variation of the dynamical environment even if its variation cannot be observed. In addition, the state space of the environment is divided evolutionarily. So, one need not to divide the state space in advance. The efficacy of the proposed system is shown by mobile robot control simulation under the environment belongs to dynamical POMDPs. The environment is the passage that has gates iterate opening and closing.

キーワード：強化学習，POMDPs，動的環境，確率的傾斜法，自律移動ロボット

Keywords: reinforcement learning, POMDPs, dynamical environment, stochastic gradient ascent method, mobile robot

1. はじめに

近年，部分観測マルコフ決定過程 (Partially Observable Markov Decision Processes: POMDPs) 環境下での強化学習の研究が盛んに行われている。実際，有効な学習方式が複数提案されてきており⁽¹⁾⁽²⁾，単純な POMDPs 環境において強化学習が行われた例は多く存在する^{(1)~(8)}。

しかし，これまで提案されている強化学習システムはシステムを構成する学習パラメータが時間的に不変であり，変化する環境で適切な政策を獲得することは難しい。そこで本論文では，その変化が観測できない動的環境という従来法では学習が困難な POMDPs 環境を挙げ，その環境においても学習が可能な手法を提案する⁽⁹⁾⁽¹⁰⁾。提案法は学習可能な時変パラメータを持つため，このような環境にも容易に適応できる。更に提案法では，GA を用いた新しい状態分割法を用いる。本手法では状態の分割が進化的に獲得されるため，設計者が強化学習システムの状態分割法をあらかじめ決定する必要がない。シミュレーションによる自律移動ロボットの行動学習により，提案法は変化が観測

できない動的環境に対応できることを示す。

2. 時変パラメータを持つ進化的強化学習システム

2.1 ルール 提案法は複数のルールにより構成される (図 1)。各ルールの前件部はそのルールが担当する環境状態を決定し，後件部は行動決定に用いるパラメータを指定する。ルールは次式により記述される。

$$R^k: \text{ if } (x_1 \text{ is } B_1^k, x_2 \text{ is } B_2^k, \dots, x_n \text{ is } B_n^k) \\ \text{ then } p^k = (p_1^k, p_2^k, \dots, p_m^k) \\ \dots\dots\dots (1)$$

ここで， $x_1, x_2, \dots, x_n$ は環境の状態ベクトル x の要素， $B_1^k, B_2^k, \dots, B_n^k$ は GA の遺伝子により決定されるクリスプ集合，そして p^k は行動決定に用いられるパラメータベクトルであり，強化学習により更新される学習パラメータである。提案法では確率的傾斜法 (Stochastic Gradient Ascent method: SGA)⁽¹⁾ を用いて学習する。

図 1 にルールが重なっている部分が存在するが，これが提案法の鍵であり，その変化が観測できない動的環境への対応を可能としている。従来のルックアップテーブルを用いた手法ではルールの重なりは存在せず，Radial Basis Function Network (RBFN)，ファジィなどを用いた手法では，ルール (RBFN では基底関数がルールに相当) の重なりは存在するが，それは単に前件部の空白を埋めるためであり，積極的に重なりを利用している訳ではない。2.2 で述べるように，提案法には時変パラメータが存在するた

* 山口大学大学院理工学研究科
 〒 755-8611 山口県宇部市常盤台 2-16-1
 Graduate School of Science and Eng., Yamaguchi University
 Tokiwadai 2-16-1, Ube-shi, Yamaguchi 755-8611
 ** 山口大学工学部
 〒 755-8611 山口県宇部市常盤台 2-16-1
 Faculty of Engineering, Yamaguchi University
 Tokiwadai 2-16-1, Ube-shi, Yamaguchi 755-8611

図 1 提案法の構成

Fig. 1. The proposed system.

め、重なりを持つルール間で各ルールの役割が時間的に変化する。そのため、提案法ではルールの重なりが積極的な意味を持つ。

2.2 行動選択 次式の重み付け総和演算により行動決定用パラメータベクトル p を決定する。

$$p_j = \frac{\sum_{k \in A} w^k(t) p_j^k}{\sum_{k \in A} w^k(t)} \dots \dots \dots (2)$$

$$w^k(t) = (\sin(w_1^k t + w_2^k) + 1) / 2 \dots \dots \dots (3)$$

ここで、 p_j はベクトル p の j 番目の要素、 A は前件部を満たすルールの集合、 $w^k(t)$ は本論文で提案する時変パラメータでありその値は $0 \leq w^k(t) \leq 1$ で変化する。そして w_1^k, w_2^k は学習パラメータである。提案法は時変パラメータを用いているため、ルールの影響は時間的に変化する。したがって、各ルールが時間を分担することが可能となり、提案法が環境の時間的な変化に対応することが可能となる。また異なる見方をすれば、提案法ではエージェントの行動は正弦波の重ね合わせで表現されるとも言える。なお、もし A が空集合となる状況や (2) 式の分母が零となる状況が生じた場合には、 p を求めず行動をランダムに決定するなど、その環境に応じた処理が必要である。また、前節で述べたルールの重なりが存在しない場合、時変パラメータの影響はなくなり、提案法は時変パラメータを持たない従来のルックアップテーブルを用いた手法と等価になる。

なお、提案法では任意の確率的行動選択法を用いることができるが、ここでは一般的によく用いられる Boltzmann 分布に基づくルーレット選択を用いる。すなわち、状態 x においてある行動 a_j が選択される確率は次式により計算される。

$$P(a_j|x) = \frac{\exp(p_j/T)}{\sum_{j=1}^m \exp(p_j/T)} \dots \dots \dots (4)$$

ここで、 T は温度定数である。なお、本選択法では行動は非連続値となるが、Adaptive PDF⁽¹¹⁾ などを用いることにより、提案法では連続値行動も容易に扱うことができる。

2.3 進化的状態分割法 提案法では、ルールの前

図 2 遺伝子の一例

Fig. 2. An example of the gene.

図 3 デコードされた領域

Fig. 3. A decoded region.

件部は遺伝子により決定され、GA により進化を行う。遺伝子の一例を図 2 に示す。ルール前件部のそれぞれのクリスプ集合に対して、同一の長さを持つストリングがそれぞれ割り当てられる。‘0’は環境状態の下半領域、‘1’は上半領域に対応する。後方のビットが担当する領域の大きさは、前方のビットの 1/2 である。そして ‘e’ は終端記号であり、これ以後のビットは無視される。たとえば、状態空間が 2 次元であると仮定すると、図 2 の 1 つ目のルールの前件部 (B_1^1, B_2^1) は図 3 のようにデコードされる。また、遺伝子の最後尾のストリング (Rule availability) はルールの有効・無効を決定する。各ルールは 1 ビットで表現され、‘0’は無効、‘1’は有効を意味する。このストリングにより、提案法では適応的にルール数を変化させることが可能である。

遺伝子 1 つは 1 個体 (1 エージェント) に対応し、GA により進化を行う。前件部のビットは、突然変異により反転又は ‘e’ へと (‘e’ は ‘0’ 又は ‘1’ へと) 変化する。

2.4 学習法 提案法では、POMDPs に対応可能な強化学習法である SGA アルゴリズムの一般形⁽¹⁾⁽⁶⁾を用いて学習パラメータの更新を行う。その SGA アルゴリズムを図 4 に示す。本研究が対象とする環境は一種の POMDPs であり、従来の actor-critic や Q-learning 等の学習法では学習が困難であると考えられる。

1. Observe $\mathbf{x}_t$ in the environment.
2. Execute action $\mathbf{a}_t$ with probability $\pi(\mathbf{a}_t, \mathbf{w}, \mathbf{x}_t)$.
3. Receive the immediate reward r_t .
4. Calculate $e_i(t)$ and $D_i(t)$ as

$$e_i(t) = \frac{\partial}{\partial w_i} \ln(\pi(\mathbf{a}_t, \mathbf{w}, \mathbf{x}_t))$$

$$D_i(t) = e_i(t) + \gamma D_i(t-1)$$
 where γ ($0 \leq \gamma < 1$) denotes the discount factor, and w_i is the i th component of $\mathbf{w}$.
5. Calculate $\Delta w_i(t)$ as

$$\Delta w_i(t) = (r_t - b) D_i(t),$$
 where b denotes the reinforcement baseline.
6. Policy Improvement: update $\mathbf{w}$ as

$$\Delta \mathbf{w}(t) = (\Delta w_1(t), \Delta w_2(t), \dots, \Delta w_i(t), \dots),$$

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha(1 - \gamma) \Delta \mathbf{w}(t),$$
 where α is a nonnegative learning rate factor.
7. Move to the time step $t+1$ and go to step 1.

図 4 SGA アルゴリズムの一般形

Fig. 4. General form of the SGA algorithm.

3. 実験

3.1 設定 Khepera Simulator version 2.0⁽¹²⁾ を用いたシミュレーションにより、エージェントがその変化を観測できない動的環境での行動学習実験を行う。その環境を図 5 に示す。図中の数値は角の座標であり、単位は mm である。中央の斜線のブロックはゲートであり、開閉を 400 ステップの周期で繰り返す。

試行について述べると、ロボットはまず図の初期状態に置かれる。エージェントはロボットの X 座標, Y 座標, 向き (角度) の 3 次元実数値ベクトルを観測し、前進, 左回転, 右回転のうち 1 つの行動を確率的に選択する。ロボットが障害物に衝突する行動を選択した場合には、罰として -1 の報酬を与え、回避動作として障害物の無い方向に回転を行う。ロボットがゴールした場合には、試行が成功したとみなし 1 の報酬を与え、その試行は終了する。また、1000 ステップが経過してもゴールに到達しないか、ロボットがゲートに挟まれる状態になった場合には、罰として -1 の報酬を与え、その試行を終了する。以上の試行の反復により、学習は進行する。なお、ロボットの現在位置のみが観測されるため、エージェントはゲートの状態の変化を観測することができない。

次に提案法の設定について述べる。ルールの各クリスプ集合のストリング長は 3 ビット, 最大ルール数は 1000 とし、遺伝子はランダムに初期化した。また GA の手法として、多様性の維持に優れ、小さい集団サイズにおいても高い性能を期待できる Minimal Generation Gap (MGG) 法⁽¹³⁾ を用いた。選択法にはルーレット選択, 交叉法には 2 点交叉を用い、集団サイズは 20 個体, そして突然変異率は (1/染色体長) とした。SGA により学習される各パラメータについては、 $p^k = 0$, w_1^k, w_2^k はランダムに初期化した。なお、GA における各個体の評価の際には 100 回の試行を行い、それらの個体の学習を行うと共に、それらの試行に

図 5 実験環境

Fig. 5. Experimental environment.

図 6 試行成功率 (提案法)

Fig. 6. Rate of success. (proposed)

図 7 試行成功率 (従来法)

Fig. 7. Rate of success. (conventional)

おけるゴール到達回数, 障害物への衝突回数, そしてルール数に基づき各個体の fitness を求める。評価対象は親 2 個体, 子 2 個体の計 4 個体である。具体的には fitness は次式により計算される。

図 8 ルール数の変化

Fig. 8. The number of rules.

図 9 Fitness の変化

Fig. 9. Fitness.

$$\text{fitness} = \xi_1 n_{\text{goal}} - \xi_2 n_{\text{collision}} - \xi_3 n_{\text{rule}} \cdots \cdots (5)$$

ここで n_{goal} , $n_{\text{collision}}$ はそれぞれ 100 試行中のゴール到達回数と衝突回数, n_{rule} は評価時のルール数である。また各 $\xi_1 \sim \xi_3$ は定数であり, $\xi_1 = 1, \xi_2 = 0.001, \xi_3 = 0.01$ とした。なお, 以上説明した本手法では, MGG 法に基づき, 学習を行った 4 個体中の 2 個体が集団に戻される。

実験では, 基本的な従来法であるルックアップテーブルを用いた SGA システム⁽⁴⁾⁽⁶⁾ (Table SGA) との比較を行う。これは図 5 においてゲートが存在しない環境では学習が行えることが確認されたシステムであり, テーブルのサイズは 8^3 , すなわち観測した 3 次元状態空間を 8^3 個のキューブにより等分割している。

3.2 結 果 提案法の結果を図 6 に示す。横軸は世代数, 縦軸は 100 試行中の成功率であり, 実線はエリート個体の結果, 破線は全 20 個体の平均である。図 7 は従来法での結果であり, 横軸は試行回数, 縦軸は過去 100 試行中の成功率である。提案法での 1000 世代は従来法での 10000 試行に相当する。これらの図が示すように, 時変パラメータを持たない従来法では, POMDPs に対応可能な学習法を用いているにもかかわらず, ゲートが存在する場合には学習が不可能であった。それに対し, 時変パラメータを持つ提案法では学習が行われることが確認された。従

図 10 ロボットの軌跡 (1~200 ステップ)

Fig. 10. Locus of a robot (1~200 step).

図 11 ロボットの軌跡 (201~400 ステップ)

Fig. 11. Locus of a robot (201~400 step).

図 12 ロボットの軌跡 (401~ゴール)

Fig. 12. Locus of a robot (401~goal).

来法では, 閉じたゲートへの衝突を繰り返すとゲートが開いている状態でもゲート位置には近づくなくなってしまうために, ゴールへの経路を獲得できなかったと考えられる。それに対し提案法は, 各ルールの役割が時間的に変化するために, 環境の時間変化への対応が可能である。次に, 提案法におけるルール数の変化を図 8 に示す。世代が進むにつれルール数が適応的に変化していることが確認できる。また, 参考までに適応度の変化を図 9 に示す。

3.3 検 討 ここでは, 提案法と従来法によりそれぞれどのような学習が成されたのか, 具体的に検討して行く。まず, 提案法により獲得された政策 (1000 世代時のエリート個体での実行結果) を図 10, 11, 12 へ示す。図 10 はロボットの 1~200 ステップ間の軌跡, 図 11 は 201~400 ステップ間の軌跡, そして図 12 は 401 ステップからゴールするまでの軌跡である。ステップ数が同一にもかかわらず, 図 11 での軌跡は図 10 の軌跡に比べ短くなっている。実際, ゲートが閉じている図 11 の状況では, ロボット

図 13 ルールによる状態分割 (初期個体)

Fig. 13. Division of the state space (initial individual).

図 14 ルールによる状態分割 (第 1000 世代時のエリート個体)

Fig. 14. Division of the state space (elite at 1000 generations).

が小刻みに左右に震えながらゆっくりと前進していることが確認されており、右側ゲートへの衝突を防ぐ政策が獲得されたものと考えられる。そしてゲートが開くとロボットの動きは再びスムーズになり、図 12 のようにゴールまで到達する。以上示したように、提案法では、時変パラメータにより環境の変化に適応した政策を獲得できる。

次に、提案法による状態分割の様子を示す。図 13 は初期個体、図 14 は第 1000 世代時のエリート個体によるものである。なお、実際には観測する状態は 3 次元 (ロボットの X 座標, Y 座標, 角度) であるが、これらの図では角度を除いた 2 次元のみを表示している。初期個体ではルールはランダムに配置されており、その重なりもランダムであるが、進化後の個体では中央部での重なりが厚くなっている。この部分は環境の時間的変化の影響が大きい部分であり、その変化に対応するためにルールの重なりが厚くなったと考えられる。なお、ルックアップテーブルを用いる従来法ではルールの重なりは存在しない。

そして次に、ゴールへ到達する政策を獲得できなかった従来法による学習の様子を示す。図 15 は過去 100 試行においてロボットが各ゲートを通じた試行の数を表し、図 16 は同じく過去 100 試行中でのロボットの平均衝突回数を表す。図 15 により、ロボットが各ゲートを通じる割合は試行を重ねることによりむしろ減少しており、Gate 2 の通過率にいたっては 0 に収束していることが確認できる。また、図 16 により、障害物への衝突回数は学習によりほぼ

図 15 ゲート通過率 (従来法)

Fig. 15. Rate of a robot passed the gate.

図 16 衝突回数 (従来法)

Fig. 16. The number of collisions.

0 にまで減少していることが確認できる。以上の結果により、従来法によりゴールへ到達する政策を獲得できなかった原因は、ロボットが Gate 2 を通過せず Gate 2 を回避するような政策を獲得してしまったためだと考えられる。すなわち、従来法では政策が時間的に固定されているため、ゲートが開いていても閉じていても同じ政策により行動を決定しなければならない。したがって、開いている Gate 2 の通過と閉じている Gate 2 の回避を両立させることができず、Gate 2 への衝突を繰り返した結果 Gate 2 自体を回避する政策が獲得されたものと考えられる。それに対し提案法では、政策を時間的に変化させることが可能であるため、たとえある時点で閉じたゲートに衝突し罰を受けたとしても、その他の時点において開いたそのゲートを通じるような学習が可能である。

4. おわりに

本章では時変パラメータを持つ進化的強化学習システムを提案し、エージェントがその変化を観測できない動的環境における学習実験を行った。POMDPs に対応可能な学習法を用いた従来法が学習できなかったこの環境においても、提案法では時変パラメータの効果により適切な行動を学習することができた。またこのような動的環境では適

切なルールの構成は難しいが、提案法ではそれが進化的に獲得されるため、設計者へ大きな負担を強いることは無い。しかし、本手法では GA を用いているため計算コストは高い。なお、実験では単純に周期的に変化する環境を扱ったが、提案法は正弦波の重ね合わせにより行動を決定するため、より複雑に変化する環境にも対応可能であると考えられる。また、提案法は時間的に変化する環境において特に有効性を発揮するが、GA により状態を分割する必要があるためオンライン処理には不向きな面を持ち、場合によっては実際の運用の前にあらかじめシステムを進化させておく必要もあると考えられる。

今後の課題としては、提案法のより詳細な解析や、複雑に変化する環境への適用、そして実問題への適用などが挙げられる。また将来的には、ルールの前件部にクリスプ集合ではなくファジィ集合を用い、状態を柔らかく分割する手法も視野に入れている。

(平成 15 年 2 月 27 日受付)

文 献

- (1) H. Kimura, M. Yamamura, and S. Kobayashi: "Reinforcement Learning in Partially Observable Markov Decision Processes: A Stochastic Gradient Method", *J. of the Jpn. Soc. for Artificial Intelligence*, Vol.11, No.5, pp.761-768 (1996-9) (In Japanese)
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」, 人工知能学誌, 11-5, pp.761-768 (1996-9)
- (2) K. Miyazaki, H. Kimura, and S. Kobayashi: "Theory and Applications of Reinforcement Learning Based on Profit Sharing", *J. of the Jpn. Soc. for Artificial Intelligence*, Vol.14, No.5, pp.800-830 (1999-9) (In Japanese)
宮崎和光・木村 元・小林重信:「Profit Sharing に基づく強化学習の理論と応用」, 人工知能学誌, 14-5, pp.800-807 (1999-9)
- (3) H. Kimura and L.P. Kaelbling: "Reinforcement Learning for Partially Observable Markov Decision Processes", *J. of the Jpn. Soc. for Artificial Intelligence*, Vol.12, No.6, pp.822-830 (1997-11) (In Japanese)
木村 元・Leslie Pack Kaelbling:「部分観測マルコフ決定過程下での強化学習」, 人工知能学誌, 12-6, pp.822-830 (1997-11)
- (4) H. Kimura, K. Miyazaki, and S. Kobayashi: "Reinforcement Learning in POMDPs with Function Approximation", *Proc. of the 14th Int. Conf. on Machine Learning*, pp.152-160 (1997-7)
- (5) K. Miyazaki, S. Arai, and S. Kobayashi: "Learning Deterministic Policies in Partially Observable Markov Decision Processes", *J. of the Jpn. Soc. for Artificial Intelligence*, Vol.14, No.1, pp.148-516 (1999-1) (In Japanese)
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」, 人工知能学誌, 14-1, pp.148-516 (1999-1)
- (6) H. Kimura and S. Kobayashi: "Reinforcement Learning for Crawling Robot Motion Using Stochastic Gradient Ascent", *J. of the Jpn. Soc. for Artificial Intelligence*, Vol.14, No.1, pp.122-130 (1999-1) (In Japanese)
木村 元・小林重信:「ロボットアームのほふく行動の強化学習: 確率的傾斜法による接近」, 人工知能学誌, 14-1, pp.122-130 (1999-1)
- (7) H. Lee, H. Kamaya, and K. Abe: "Performance of LQ-learning in POMDP Environments", *Proc. of SICE Annual Conference 2002 (SICES2002)*, pp.922-925 (2002-8)
- (8) S.A. Guirnaldo, K. Watanabe, K. Izumi, and K. Kiguchi: "Reinforcement Learning with Expectation and Action Augmented States in Partially Observable Environment", *Proc. of SICE Annual Conference 2002 (SICES2002)*, pp.926-931 (2002-8)
- (9) K. Umesako, M. Obayashi, and K. Kobayashi: "Evolutionary Reinforcement Learning System with Time-Varying Parameters", *Proc. of 2002 Int. Conf. on Control, Automation and Syst. (ICCAS2002)*, pp.1284-1287 (2002-10)
- (10) K. Umesako, M. Obayashi, and K. Kobayashi: Evolutionary and time-varying reinforcement learning system for unobservable dynamic environment, *Proceedings of the Eighth International Symposium on Artificial Life and Robotics (AROB 8th '03)*, pp.82-85 (2003-1)
- (11) K. Umesako, M. Obayashi, and K. Kobayashi: "Reinforcement Learning Using Adaptive Search Method", *IEEE Trans. EIS*, Vol.122-C, No.3 pp.374-380 (2002-3) (In Japanese)
梅迫公輔・大林正直・小林邦和:「適応的探索法を用いた強化学習」, 電学論 C, 122-3, pp.374-380 (2002-3)
- (12) "Khepera Simulator homepage", <http://diwww.epfl.ch/lami/team/michel/khep-sim/index.html>
- (13) H. Sato, I. Ono, and S. Kobayashi: "A New Generation Alternation Model of Genetic Algorithms and Its Assessment", *J. of the Jpn. Soc. for Artificial Intelligence*, Vol.12, No.5, pp.734-744 (1997-9) (In Japanese)
佐藤 浩・小野 功・小林重信:「遺伝的アルゴリズムにおける世代交代モデルの提案と評価」, 人工知能学誌, 12-5, pp.734-744 (1997-9)

梅 迫 公 輔 (非会員) 1976 年生。1998 年 3 月広島工業大学工学部電子工学科卒業。2000 年 3 月山口大学大学院理工学研究科博士前期課程知能情報システム工学専攻修了。同年 4 月同大学院理工学研究科博士後期課程システム工学専攻入学。計測自動制御学会会員。

大 林 正 直 (正員) 1951 年生。1975 年 3 月九州工業大学工学部電気工学科卒業。1977 年 3 月九州大学大学院工学研究科修士課程電気工学専攻修了。同年 4 月(株)日立製作所入社大みか工場勤務。1989 年 4 月九州大学助手, 1997 年 7 月同助教授, 1998 年 4 月山口大学助教授, 現在, 同教授。計測自動制御学会, システム制御情報学会, 米国電気電子学会, 日本神経回路学会の各会員。博士(工学)

小 林 邦 和 (非会員) 1969 年生。1991 年 3 月山口大学工学部電子工学科卒業。1993 年 3 月同大学院工学研究科博士前期課程知能情報システム工学専攻修了。大学院博士後期課程システム工学専攻退学。1994 年 4 月同大工学部知能情報システム工学科助手。現在に至る。米国電気電子学会, 電子情報通信学会, 日本神経回路学会の各会員。