リレー解説に寄せて：強化学習研究の歩み

木　村　元*

*九州大学 大学院工学研究院，福岡県福岡市西区元岡 744
*Graduate School of Engineering, Kyushu University, Motooka 744 Nishi-ku Fukuoka city 819-0395, Japan
*E-mail: kimura@nams.kyushu-u.ac.jp

キーワード：強化学習 (reinforcement learning), 分類子システム (classifier system), Q 学習 (Q-learning), 政策勾配法 (policy gradient method), actor-critic method
JL 002/02/4202–0086

1. はじめに

強化学習とは，ゴール指向型の学習と意思決定を理解し実現するための計算論的な接近法であり，試行錯誤による環境との相互作用を通じて学習していくという点において，教師による教示や環境の完全なモデルに依存する他の学習とは区別される[13]．この強化学習（reinforcement learning）の研究には，大きく3つの流れが存在したと考えられる．第一の流れとして分類子システム[5] 第二はニューラルネットワーク関連の研究における試行錯誤学習，第三にマルコフ決定過程の最適制御問題とダイナミックプログラミングが挙げられる．このほか学習オートマトンにおいても強化学習の研究がみられるが、原理的には勾配法が多いので本稿ではニューラルネットワークに含めて紹介する．それぞれの流れにおいて主流となる学習アルゴリズムや解析手法は異なるが，これらの異なる分野同士で影響を及ぼし合いながら研究が大きく発展してきた．本稿では，主に 1990 年代から近年に至るまでのおよそ 20 年間の強化学習研究におけるいくつかのエポックメイキング的な手法や解析を中心に紹介し，それらの関係を示すことにより強化学習研究の歩みについて概観する．

2. 分類子システムにおける信頼度割当

分類子システム (classifier system) とは，分類子 (classifier) と呼ばれるルールを学習することにより，与えられた環境に適応するようなプロダクションシステム，すなわち知識ベースおよび推論システムである[3]．以下の3つの機能により構成される：

- 感覚入力に応じて行動出力のルールを選択する機能
- 環境から与えられる報酬に基づいてルールに信頼度 (credit) を与える機能
- 新しいルールを生成・発見する機能

このうち 3 番目の機能は遺伝的アルゴリズムによって与えられる．本稿では強化学習を実現する上で重要な 1 番目と 2 番目の機能について紹介する．分類子システムのルールは以下のような形式である：

if ＜条件＞ then ＜行動＞

ルールの条件部は固定長の文字列で表され，環境からの感覚入力の文字列とマッチするルールが選択される．このとき，

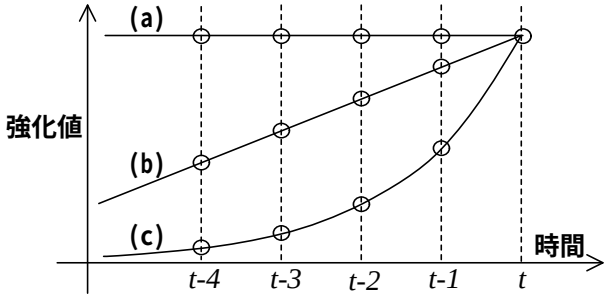

図 1　Profit sharing におけるルールの信頼度割当の例．

複数のルールがマッチする可能性があるが，その場合ルールに付加されている信頼度（重み変数）に基づいてルール選択の意思決定を行う．感覚入力と行動出力により環境中で状態遷移を繰り返し，それら一連のエピソードの結果としてエージェントは報酬を獲得する．このとき分類子システムはルールへの信頼度の割当機能により，報酬獲得に有効なルール群を獲得していく．信頼度割当 (credit assignment) のアルゴリズムとしてはまずバケツリレー法 (bucket brigade method) が提案された．これは，報酬を獲得したエピソードにおいて実行されたルールを時系列順に考え，時間的に後ろのルールから前のルールへと bid と呼ばれる一定値を伝播しながら各ルールの信頼度へ加えていく方法である[5]．時間的に隣接するルール間で値をやりとりすることから，後述する TD 法との類似性が指摘されている[13]．このバケツリレー法では，エピソードのステップ数が多い場合，エピソードの初期に実行するルール群を強化するための bid の伝播に時間がかかるという問題があった．そこで，エピソードで実行したルールとその順番を記憶しておき，環境から報酬を得た時点でそれらのルールに対して一気に信頼度割当を行う profit-sharing 法が提案された．図 1 に profit-sharing 法によるルールの信頼度割当による強化の例を示す．

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い，経験した全てのエピソード上で常に迂回系列上にあるルールを無効ルールと呼ぶ．Profit-sharing 法によるルールの強化において，任意の無効ルールが抑制されるための必要十分条件は以下で示される：

1. 【状態観測】エージェントは環境の状態 s_t を観測し，特徴ベクトル $\boldsymbol{x}(t) = (x_1(t), x_2(t), \cdots x_m(t))$ を生成する．
2. 【ASE による行動選択】行動 $a(t)$ を以下の式で決める：

$$a(t) = f\left(\sum_{i=1}^{m} \theta_i x_i(t) + \mathrm{noise}(t)\right)$$

ただし θ_i は ASE の重み変数を表し，この値を学習によって変更していく．関数 f は以下のようなしきい値関数である：

$$f(x) = \begin{cases} 1 & , \text{if } x \geq 0 \\ -1 & , \text{if } x < 0 \end{cases}$$

また noise は期待値 0 分散 σ^2 の正規分布である．
3. 【ACE による行動評価】状態遷移後の状態 s_{t+1} を観測し，特徴ベクトル $\boldsymbol{x}(t+1)$ を生成する．以下の式に従って状態価値関数 $V(t)$ および $V(t+1)$ を計算する：

$$V(t) = f\left(\sum_{i=1}^{m} w_i x_i(t)\right)$$

この状態評価値を利用して行動評価値 TD_error を計算する：

$$\mathrm{TD\_error} = r(t) + \gamma V(t) - V(t-1)$$

ただし γ は割引率，r_t は状態遷移に伴って得た報酬である．
4. 【行動の強化】ASE の重み変数 θ_i を以下のように更新：

$$\theta_i \leftarrow \theta_i + \alpha \mathrm{TD\_error}\, e_i(t)$$

ただし $\alpha > 0$ は学習率，$e_i(t)$ は ASE の適正度の履歴 (eligibility trace) で以下のように計算する：

$$e_i(t) = \delta e_i(t-1) + (1-\delta) a(t) x_i(t)$$

ただし $0 \leq \delta < 1$ は減衰率である．
5. 【ACE の更新】以下のように ACE の重み変数 v_i を更新：

$$w_i \leftarrow w_i + \beta \mathrm{TD\_error}\, \overline{x_i}(t)$$

ただし $\beta > 0$ は学習率，$\overline{x_i}$ は ACE の適正度の履歴 (eligibility trace) で以下のように計算する：

$$\overline{x_i}(t) = \lambda \overline{x_i}(t-1) + (1-\lambda) x_i(t)$$

ただし $0 \leq \lambda < 1$ は減衰率である．
6. 時間ステップ t を $t+1$ へ進めて手順 1 へ戻る．

図 2　Barto らの ASE-ACE アルゴリズム

$$L \sum_{i=1}^{W} f_{t-i} < f_t\,, \quad \text{for all } t = 1, 2, \cdots W-1.$$

ここで W はエピソード長，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．これは無効ルール抑制定理[11] と呼ばれ，この条件を満たす強化関数の例としては図 1(c) のような指数関数がある．

分類子システムにおける解析では，上記のように環境のマルコフ性を仮定せず，また利得の最適性ではなく無効ルールの抑制に着目するなどの点でユニークな特徴があり，後に紹介する強化学習のアプローチとは一線を画する．

3. ニューラルネットワークにおける試行錯誤学習と政策勾配法

生体の神経回路網についての知見から着想を得て人工神経回路網モデルを構築し解析することにより生体の学習について理解し，工学的に応用していくというのがニューラルネットワーク研究における主要なスタンスである．この流れから Hebb 則やパーセプトロン，バックプロパゲーションなどの目覚しい成果が生まれたのは周知のことであるが，当然ながら試行錯誤学習についても同様のアプローチが試みられた．Barto らはパーセプトロンに類似した人工神経素子とノイズ発生器を組み合わせた試行錯誤による行動選択器 ASE(Associative Search Element) と，報酬の予測学習器 ACE(Adaptive Critic Element) を組み合わせ，これらを Hebb 則によって学習させることにより強化学習システムを構築し，倒立振子のコントローラの学習への適用を示した[1]．図 2 はその ASE-ACE 学習器のアルゴリズムを示す．他の関連するアルゴリズムとの関係を分かりやすく示すために変数等の記法をオリジナルから多少変更している．また，図 2 のアルゴリズム中では状態観測 s_t が特徴量 $\boldsymbol{x}(t)$ に変換されているが，論文中では倒立振子の台車の位置と速度および振子の角度および角速度の 4 次元連続値の状態量を $3 \times 3 \times 6 \times 3 = 162$ の領域に分割しており，特徴ベクトルの要素数 $m = 162$ である．行動出力は $+1$ または -1 の 2 値である．文献中においてアルゴリズムの妥当性について解析は示されていないものの，後に actor-critic 法と呼ばれるアルゴリズム群の基本的なアイデアがほぼ全て含まれている．ASE と呼ばれる部分が actor に該当し，ACE が critic に相当する．特に ACE 部分の学習アルゴリズムは TD(λ) 法と呼ばれている．共著者の一人の Sutton は，ACE の処理に注目し，環境がマルコフ過程である場合において TD(λ) 法が割引報酬の期待値を推定していることを解析により示し，収束を証明した[13]．これにより TD(λ) の $\lambda = 1$ の場合はモンテカルロ法による割引報酬のサンプル平均になり，$\lambda = 0$ の場合はマルコフ過程における非同期ダイナミックプログラミングとなり，λ がその中間にある場合はそれら双方の性質を兼ね備え，学習速度の点でより効果的であることを指摘し，適正度の履歴 (eligibility trace) の意味について初めて理論的な解釈を与えた．

行動選択器 ASE において，その更新の手がかりとしている行動評価値 TD_error は，この呼び方は本稿の著者が都合によりこの記法で示したが，オリジナルの論文中では直接報酬の推定値 $\hat{r}$ となっている．Barto らが ACE を併用してこの TD_error を手がかりに学習する手法を提案する以前は，TD_error= $\hat{r}$ の代わりに直接報酬 r_t を用いて学習する手法が存在しており，その方法でもある程度学習できることが示されていた．これは ASE(actor) の行動選

択法と ASE 中の適正度の履歴を用いた学習アルゴリズムでも学習できることを意味し，これに密接に関連するアルゴリズムが episodic REINFORCE アルゴリズムとして示されている[16]．これは，エージェントが時刻 t の状態 s_t において行動 a_t を確率（あるいは確率密度）$g(a_t, \boldsymbol{\theta}, s_t)$ に従って選択することを k ステップの期間行い，その状態-行動系列の結果として利得 V が与えられた場合，エージェントの重み変数ベクトル $\boldsymbol{\theta}$ を以下のように更新する：

$$\Delta\boldsymbol{\theta} = \alpha\,(V - b)\sum_{t=1}^{k} \nabla_{\boldsymbol{\theta}} \ln g(a_t, \boldsymbol{\theta}, s_t)$$

ただし $\alpha > 0$ は学習率，b は reinforcement baseline と呼ばれる定数パラメータである．報酬 r の期待値がエージェントの重み変数 $\boldsymbol{\theta}$ の条件付期待値で与えられる場合，

$$\mathrm{E}\{\Delta\boldsymbol{\theta}\} = \alpha \nabla_{\boldsymbol{\theta}} \mathrm{E}\{V|\boldsymbol{\theta}\}$$

となることが証明されている．また b の値は利得 V の期待値の値と同じになると $\Delta\boldsymbol{\theta}$ による更新のランダムウォークの度合いが最も低減され，確実に利得の期待値の 1 次勾配を登るように重み変数を更新するようになる．ここで $g(a_t, \boldsymbol{\theta}, s_t)$ は確率的政策であり $\nabla_{\boldsymbol{\theta}} \ln g(a_t, \boldsymbol{\theta}, s_t)$ はその適正度 (eligibility) である．利得 V を割引報酬として各ステップで報酬が入るものとしてアルゴリズムを構築すると，ASE(actor) の行動選択法と ASE 中の適正度の履歴を用いた学習アルゴリズムが導かれる．また，図 2 のアルゴリズムのように TD_error と ASE 中の適正度の履歴を用いて actor を更新する方法は，先の episodic REINFORCE アルゴリズムにおける b の部分を利得 V の推定値としていることと等価であることが示されている[8]．

このように actor の行動選択のメカニズムを行動の確率分布関数（確率的政策）に従って駆動し，利得の勾配方向へ政策のパラメータを更新する方法は「政策勾配法」と総称され，行動が連続値でも離散値でもアルゴリズムが比較的簡単に実装可能なのでロボットの学習制御などに実績がある[18]．さらに，actor の政策を改善するのに必要十分な価値関数近似のための基底関数は何なのかについての研究が進められ，政策勾配定理[14] により、actor の政策を改善するのに必要な価値関数近似のための基底関数は、適正度すなわち行動選択確率関数（政策）の対数を政策パラメータで偏微分した関数であることが示されている。確率的政策 π は，状態 s において行動選択確率 $\pi(s, a)$ に従って行動 a を実行する．この $\pi(s, a)$ は政策パラメータ群 $\boldsymbol{\theta}$ によって値を調節でき，パラメータ群の各要素 θ_i により偏微分可能な関数であるとする．ここで利得 V を以下で定義する：

$$V^{\pi}(s) = E\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| s_0 = s, \pi\right] . \qquad (1)$$

エージェントの学習目標である目的関数 $\rho(\pi)$ を定義する：

$$\rho(\pi) = V^{\pi}(s_0), \quad Q^{\pi}(s, a) = E\{V_t | s_t = s, a_t = a, \pi\}$$

ここで $d^{\pi}(s) = \sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s|s_0, \pi)$ とすると，政策勾配定理[14] より以下の式が与えられる：

$$\frac{\partial \rho}{\partial \theta} = \sum_{s\in\mathcal{S}} d^{\pi}(s) \sum_{a\in\mathcal{A}} \frac{\partial \pi(s, a)}{\partial \theta} Q^{\pi}(s, a) . \qquad (2)$$

次に，真の Q 関数 $Q^{\pi}(s, a)$ を関数近似によって $\hat{Q}^{\pi}(s, a)$ と表すことを考える。ただしこの近似 Q 関数 $\hat{Q}^{\pi}(s, a)$ は適正度 $\frac{\partial \ln \pi(s,a)}{\partial \theta_i}$ を各要素としたベクトル $\nabla_{\theta} \ln \pi(s, a)$ を基底関数とし、この各適正度に対応した要素を持つパラメータベクトル $\boldsymbol{w}$ との線形結合および状態価値関数 $V^{\pi}(s)$ を用いて以下のように表す：

$$\hat{Q}^{\pi}(s, a) = (\nabla_{\boldsymbol{\theta}} \ln \pi(s, a))^T \boldsymbol{w} + V^{\pi}(s) . \qquad (3)$$

真の Q 関数 $Q^{\pi}(s, a)$ と近似 Q 関数を $\hat{Q}^{\pi}(s, a)$ との 2 乗誤差が最小になるようパラメータベクトル $\boldsymbol{w}$ を調節すると，政策勾配定理[14] より以下が成り立つ：

$$\frac{\partial \rho}{\partial \theta} = \sum_{s\in\mathcal{S}} d^{\pi}(s) \sum_{a\in\mathcal{A}} \frac{\partial \pi(s, a)}{\partial \theta} \hat{Q}^{\pi}(s, a) . \qquad (4)$$

つまり、critic において Q 関数を近似する場合に actor の適正度を基底関数として利用すると，Q 関数が近似であるにもかかわらず真の勾配方向へ政策を更新できることを意味する．一般に Q 関数の表現では，以下のように状態評価値 $V^{\pi}(s)$ と，advantage と呼ばれる評価値 $A^{\pi}(s, a)$ を用いて $Q^{\pi}(s, a) = V^{\pi}(s) + A^{\pi}(s, a)$ のように表すことができる．式 (3) の右辺左側の項 $(\nabla_{\theta} \ln \pi(s, a))^T \boldsymbol{w}$ がこの advantage 関数 $A^{\pi}(s, a)$ を近似している．

この式 (4) の 1 次偏導関数で示される政策勾配を使った様々な政策改善アルゴリズム[10] が提案される一方で，2 次偏導関数も考慮に入れた「自然勾配 (natural gradient)」と呼ばれる勾配を用いる政策改善法が提案されている[7]．この種の勾配法は，1 次偏導関数しか使用しない勾配法に比べて格段に少ない反復回数で極値を得ることができ，一般的な最適化手法としてニュートン法が知られているが，一般に 2 次偏導関数行列（ヘッセ行列）の逆行列を求めるという大きな計算コストが要求される．ところが強化学習における政策パラメータに関する自然勾配は，情報理論的に興味深い性質を持ち、上記の逆行列計算が不要になる．Advantage 関数を構成する式（3）のパラメータベクトル $\boldsymbol{w}$ において，ある政策パラメータ θ の適正度に対応する要素を w_{θ} と表す．これらが式（4）を満たすとき，目的関数 $\rho(\pi)$ に対するパラメータ θ の自然勾配 $\frac{\partial \tilde{\rho}}{\partial \theta}$ は，以下のように驚くべき単純な式で表される[12]：

$$\frac{\partial \tilde{\rho}}{\partial \theta} = w_{\theta} . \qquad (5)$$

すなわち，基底関数として actor の適正度を用いた advantage 関数を構成するパラメータベクトル $\boldsymbol{w}$ は，対応する

政策パラメータ $\boldsymbol{\theta}$ の自然勾配になっている。このアルゴリズムは、Q 値を推定しつつ２次勾配も考慮しながら政策を改善していくため，マルコフ性を満たす大規模問題において極めて効率良く政策改善を行うことが期待される．

4. マルコフ決定過程の最適制御問題とダイナミックプログラミング

マルコフ決定過程 (Markov Decision Process: MDP) における最適制御問題において Bellman の最適方程式と呼ばれる関数方程式を立て，この方程式を解いて解を得る手法のクラスは動的計画法 (dynamic programming: DP) として知られる．環境をマルコフ決定過程でモデル化し，DP で最適政策を得ることはプランニングと呼ばれ，完全な環境モデルを要することから強化学習とは区別されるが，DP の計算における状態遷移確率による重み付けの部分を確率サンプリングによる平均操作で代用し，それを未知の環境中でエージェントが試行錯誤を行うことにより実現した強化学習法が Q-learning[15] である．さらに Q 値などを関数近似で表現した場合におけるアルゴリズムや収束についての詳しい解析が示された[2]．

5. おわりに

本稿では主に３つの系統の強化学習アルゴリズム研究の歩みについて概説した．この他の重要な強化学習の話題として以下が挙げられる：

- **連続・多次元空間への対応：**状態空間については関数近似法[2] や特徴ベクトルの生成問題として扱われる．空間の適応的分割などの手法[17][4] もこれに相当する．行動空間については特に行動選択方法が問題になる．その場合，確率密度関数を確率的政策として連続値行動を扱い，政策勾配法を適用するのが一般的[18] であるが，ランダムタイリングと Gibbs-sampling を組合わせて多次元状態-行動空間で Q-learning を行う方法[9] がある．
- **状態観測の不完全性への対応：**マルコフ決定過程では，状態観測が完全であることが前提であるが，実環境や応用においては完全な状態観測が困難である場合がある．これをモデル化したのが部分観測マルコフ決定過程 (partially observable MDP: POMDP) であり，この環境で学習するための方法が研究されている．
- **マルチエージェント：**マルコフゲームでの学習[6] など．

(2012 年 9 月 21 日受付)

参 考 文 献

1) Barto, A. G., Sutton, R. S. & Anderson, C. W.: Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems, *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-13, no.5, September/October 1983, pp.834–846.

2) Bertsekas, D.P. and Tsitsiklis, J.N.: "Neuro-Dynamic Programming", Athena Scientific (1996).

3) Goldberg, D. E.: Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley, Reading,MA (1989).

4) 半田 久志，二宮 明，堀内 匡，小西 忠孝，馬場 充：強化学習における矛盾の概念に沿った漸増的な状態空間の構成法，計測自動制御学会論文集 Vol.38, No.5, pp.469–476 (2002).

5) Holland, J. H.: Escaping brittleness, Machine Learning, an artificial intelligence approach. Volume II. R. S. Michalski, J. G. Carbonell and T. M. Mitchell ed., Morgan Kaufmann, pp. 593–623 (1986).

6) Hu, J. & Wellman, M. P.: Multiagent Reinforcement Learning: Teoretical Framework and an Algorithm, Proceedings of the 15th International Conference on Machine Learning, pp. 242–250 (1998).

7) Kakade, S.: A Natural Policy Gradient, Advances in Neural Information Processing Systems 14, pp.1531–1538 (2002).

8) 木村 元，小林 重信: Actor に適正度の履歴を用いた Actor-Critic アルゴリズム– 不完全な Value-Function のもとでの強化学習，人工知能学会誌, Vol.15, No.2, pp.267–275 (2000).

9) 木村 元：ランダムタイリングと Gibbs-sampling を用いた多次元状態-行動空間における強化学習，計測自動制御学会論文集, Vol.42, no.12, pp.1336–1343 (2006).

10) Konda, V.R. & Tsitsiklis, J.N.: Actor-Critic Algorithms, Advances in Neural Information Processing Systems 12, pp. 1008–1014 (2000).

11) 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.811–821 (1997).

12) Peters, J., Vijayakumar, S. & Schaal, S.: Reinforcement Learning for Humanoid Robots - policy gradients and beyond, 3rd IEEE-RAS International Conference on Humanoid Robotics (2003).

13) Sutton, R. S. and Barto, A.: "Reinforcement Learning: An Introduction", A Bradford Book, The MIT Press (1998).

14) Sutton, R. S., McAllester, D., Singh, S. & Mansour, Y.: Policy Gradient Methods for Reinforcement Learning with Function Approximation, Advances in Neural Information Processing Systems 12 (NIPS12), pp. 1057–1063 (2000).

15) Watkins, C.J.C.H. & Dayan, P.: Technical Note: Q-Learning, Machine Learning 8, pp.279–292 (1992).

16) Williams, R. J.: Simple Statistical Gradient Following Algorithms for Connectionist Reinforcement Learning, *Machine Learning 8*, pp. 229–256 (1992).

17) 矢入 健久，堀 浩一，中須賀 真一: 複数行動結果を考慮した最尤推定に基づく状態一般化法，人工知能学会誌, Vol.16, No.1, pp.128–138 (2001).

18) 吉本 潤一郎，銅谷 賢治，石井 信：強化学習の基礎理論と応用，計測と制御, Vol.44, No.5, pp.313–318 (2005).

［著 者 紹 介］

木村 元

1992 年東京工業大学工学部制御工学科卒業．1997 年同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．1998 年 4 月，同大学大学院総合理工学研究科助手．2004 年 4 月，九州大学大学院工学研究院海洋システム工学部門助教授，2007 年 4 月，同部門において准教授へ職名変更，現在に至る．情報技術の船舶海洋分野への応用に関する研究に従事．計測自動制御学会，日本船舶海洋工学会，人工知能学会，日本ロボット学会会員．