原著論文

積極性と警戒心のバイアスを 加味した強化学習†

瀧田 航一朗 萩原 将文*

本論文では、動物における積極性と警戒心に対応するような形で価値判断の変更を行う、新しい強化学習モデルを提案する。強化学習は、教師あり学習に比べより幅広い状況において利用できる学習方式ではあるが、環境からエージェントに与える強化信号の量を適切に定めることは、必ずしも容易ではない。例えば、次善の行動に過度な報酬を与えてしまえば、最善の行動の探索が阻害されうるし、逆に報酬が不足であれば、次善の行動すらも学習されないということが起こりうる。このような問題に対し、本提案モデルでは、報酬の最大化のみを目的とした、積極性に対応するモジュールと、罰の最小化のみを目的とする、警戒心に対応したモジュールとを組み合わせ、そのバイアスを調整することで、幅広い環境への適応を図る。計算機実験により、従来モデルでは性能が極端に悪化するような環境においても、提案モデルが優れた性能を示すことを確認した。また、それ以外の環境においても、提案モデルは従来モデルと同等の性能を示した。

キーワード：強化学習, actor-critic, 情動

1 はじめに

近年、社会への普及度が高まりつつある介護・娯楽・案内ロボットなどにおいて、学習に基づく行動制御は今後その重要性を増していくと考えられる。工場などの限定的かつ既知の環境とは違い、一般の住宅内などでは常に不測の事態を想定せねばならず、学習による汎化性と、環境変化への適応能力は必須のものとなるからである。

特に強化学習[1]に基づくアルゴリズムは、教師あり学習に基づくものよりも広範な利用が可能である。そもそも強化学習とは、環境のダイナミクスが与えられていない状況下において、trial-and-error に基づいて受け取る強化信号の最大化を図る学習方式である。教師あり学習では、各状況下における正しい出力そのものを外部から教える必要があるのに対し、強化学習においては、エージェントの出力の望ましきの度合だけを強化信号の形で与えてやれば良い。例えば、ロボットがユーザの役に立った場合にはロボットの頭を撫でてやる、逆の場合には頭を叩くというようにすれば、エキスパートの手を必要とせずに、ユーザにとって望ましい行動を学習させることができる。近年、特に部分観測環境における学習や、マルチエージェント環境

における強化学習の研究が進み[2]～[5]、多くの研究者の注目を集めている。

強化学習アルゴリズムを用いるに際して、報酬(正の強化信号)と罰(負の強化信号)の量をどのように設定するかは、学習結果を大きく左右する重要な問題である。例えば、必ずしも最善ではない行動に過大な報酬を与えた場合には、最善の行動の探索が阻害されうる。逆に過小であれば、罰信号に強く影響されて消極的な行動しか選択されなくなってしまうことがある。このような問題は特に複数の報酬源が存在する場合や、回避不可能な罰信号の後に報酬が与えられるような環境で顕在化する。

一方、実際の生物がどのように報酬や罰に対して反応するかということを考えてみると、特に哺乳類などの高等な生物においては、様々な情動が大きな役割を果たしているということが出来る。動物実験において、同じ報酬を与えた場合にも、報酬に対する期待の度合が違ふと異なった反応を示す事が知られている[6], [7]。また、怒りや警戒心といった情動に支配されている動物は、飼い主や餌といった本来好ましい対象に対して、普段と異なる反応を示す場合が見られる。これは、長期的な学習とは別に、一時的かつ動的な価値判断の変更を行うことで、環境への適応を図っているものと考えられる。仮に、周囲に危険が存在することが既知である場合には、より詳しい状況が分かるまでは餌を獲得するよりも危険回避を優先する方が重要であるし、逆に周囲の環境が安全かつ安定であると推測されれば、より良い餌場を求めて探索行動をする方が合理的たりうる。

† Reinforcement Learning Algorithm with Bias of Activeness and Canutiousness

Koichiro TAKITA and Masafumi HAGIWARA

* 慶應義塾大学
Keio University

以上のような背景から、本論文では、動物における積極性と警戒心に対応するような形で価値判断の変更を行う強化学習手法を提案する。具体的には、報酬の最大化のみを目的とした、積極性に対応するモジュールと、罰の最小化のみを目的とした、警戒心に対応するモジュールで構成され、出力に対する各モジュールのバイアスを調整することにより学習の効率化を行う。

提案モデルの構造としては、actor-critic モデル[8]～[10]を元にしたものを採用した。actor-critic モデルは、状態評価を行う critic モジュールと行動選択を行う actor モジュールよりなるモデルで、環境同定型の強化学習手法の中でも政策反復法(policy iteration)に分類される。

一般にマルコフ決定過程においては、同じ環境同定法でも価値反復法(value iteration)に分類される Q-Learning[1], [11]が非常に有効であるとされ、広く用いられているが、マルコフ性が成立しない環境やマルチエージェント環境においては、actor-critic モデルが有効である。また actor-critic モデルは、Q-Learning と比べ、連続値の出力が要求されるような環境への応用が非常に容易であるという長所を持つ[1]。

また、経験強化型の手法の中では、行動系列(エピソード)単位で学習を行い、獲得した報酬を系列中の各行動に分配する、Profit Sharing 法[12], [13]などが広く用いられている。Profit Sharing 法は学習途中においても効率的に報酬を獲得し続けるという点で非常に優れているものの、最適解ではなく準最適解に陥る危険がある。

以上のような各手法の特徴を踏まえ、提案モデルでは、応用性の高さと学習の安定性を重視し、actor-critic モデルを元にした構造を採用した。

計算機シミュレーションにより、提案モデルが、一般の環境においては従来モデルと同等の性能を保ちつつ、回避不可能な罰信号の後に報酬が与えられるような環境においては、従来モデルを大きく上回る性能を発揮することを確認した。

2 では、従来型の actor-critic モデルの一例を簡潔に説明する。3 では、提案モデルについて述べる。4 では、従来型のモデルと提案モデルとを比較した。

2 Actor-Critic モデル

図 1 に、一般的な actor-critic モデル[1]の構成を示す。エージェントは状態評価を行う critic モジュールと行動選択を行う actor モジュールの二つのモジュールから構成される。

エージェントが時刻 t において環境から状態 s_t を観測すると、actor は内部の行動評価値に基づき行動 a_t を実

行する。その結果として、エージェントは環境から強化信号 r_t を受け取り、次の状態 s_{t+1} を観測する。Critic モジュールは、 s_t から s_{t+1} への状態遷移の評価値である TD 誤差 δ_{t+1} を

$$\delta_{t+1} = r_t + \gamma V_t(s_{t+1}) - V_t(s_t) \quad (1)$$

に基づき計算する。ここで、 γ は未来の報酬に対する割引率であり、 $V_t(s_t)$ は状態 s_t に対する評価値である。

TD 誤差が正である場合、実行した行動 a_t は望ましいものであったと考えられ、逆に TD 誤差が負の場合には、 a_t は望ましくない行動であったと考えられる。そこで、actor は、critic から受け取った TD 誤差を元に、状態 s_t における行動 a_t の評価値 $p_t(s_t, a_t)$ を

$$p_{t+1}(s_t, a_t) = p_t(s_t, a_t) + \alpha \delta_t \quad (2)$$

のように更新する。

また、critic は状態 s_t に対する評価値 $V_t(s_t)$ を

$$V_{t+1}(s_t) = V_t(s_t) + \beta \delta_t \quad (3)$$

として更新する。なお、 α 、 β はそれぞれ、actor と critic の学習係数である。

Actor モジュールの行動選択法は無数に考えられるが、広く用いられている方法として、

$$\pi_t(s_t, a) = \frac{\exp(p_t(s_t, a)/T)}{\sum_a \exp(p_t(s_t, a')/T)} \quad (4)$$

で定義される Boltzman 選択がある。ここで、 $\pi_t(s_t, a)$ は状態 s_t において行動 a を選択する確率、 $p_t(s_t, a)$ は状態 s_t に対する行動 a の actor による評価値、 T は行動選択のための温度パラメータである。

図 1 Actor-critic モデル

Fig.1 Actor-critic model.

3 積極性と警戒心のバイアスを加味した強化学習

前述の従来型 actor-critic モデルに対し、提案モデルは、図 2 に示されるような構造を持つ。この構造の特徴は、正の強化信号である報酬のみに基づいて動作する actor と critic の組と、負の強化信号である罰のみに基づいて動作する actor と critic の組、計 4 つのモジュールから構成されているという点にある。ここでは、報酬のみに基づいて動作する actor と critic を積極的モジュールと呼び、罰のみに基づいて動作する actor と critic を警戒的モジュールと呼ぶことにする。

また、エージェントはこれらモジュールが行動選択に及ぼす影響を左右する、バイアスパラメータを持つ。初期状態では、エージェントの行動は、警戒的モジュールと積極的モジュールの両方から均等な影響を受けて決定される。しかし、バイアスパラメータが上昇すると、警戒的モジュールの比重が減っていき、積極的な行動が優先されるようになる。バイアスパラメータは、後述するように、罰を受けている場合には減っていき、報酬を得ている場合には増えていくという傾向を持つ。

これにより、新しい環境ではまず罰の回避と、少量でも安定した報酬の獲得が達成される。そして、環境に適応しはじめると多少のリスクを負ってもより大きな報酬の探索を目指すようになる。つまり、実際の動物で考えれば、未知の環境では短期的な生存本能に従った行動、たとえば敵の回避や飢えの克服がまず優先され、安全が確保された後で初めて、より大きな獲物を狙うようになるわけである。

具体的な行動選択と学習の手順を以下に示す。エージェントが時刻 t において環境から状態 s_t を観測すると、積極的 actor と警戒的 actor は、バイアスパラメータ e_t と内部の行動評価値とに基づき行動 a_t を実行する。

図 2 提案モデル
Fig.2 Proposed model.

状態 s_t における行動 a_t の、積極的 actor にとっての評価値を $p_t^+(s_t, a_t)$ とし、警戒的 actor にとっての評価値を $p_t^-(s_t, a_t)$ とすると、総合的な評価値 $p_t(s_t, a_t, e_t)$ を、

$$p_t(s_t, a_t, e_t) = p_t^+(s_t, a_t) \cdot e_t + p_t^-(s_t, a_t) \cdot (1 - e_t) \quad (5)$$

と定める。ただし、バイアスパラメータ e_t については、 $0 \leq e_t \leq 1$ が成り立つものとする。

行動選択法としては、

$$\pi_t(s_t, a, e_t) = \frac{\exp(p_t(s_t, a, e_t)/T)}{\sum_a \exp(p_t(s_t, a', e_t)/T)} \quad (6)$$

で定義される Boltzman 選択などを用いる。ここで、 $\pi_t(s_t, a, e_t)$ は状態 s_t において行動 a を選択する確率、 $p_t(s_t, a, e_t)$ は前述の総合的な評価値、 T は行動選択のための温度パラメータである。

以上のようにして選択された行動の結果として、エージェントは環境から強化信号 r_t を受け取り、次の状態 s_{t+1} を観察する。積極的 critic モジュールは、 s_{t+1} への状態遷移に対して、報酬に基づく TD 誤差 δ_{t+1}^+ を

$$\delta_{t+1}^+ = \begin{cases} r_t + \gamma V_t^+(s_{t+1}) - V_t^+(s_t), & r_t > 0 \\ \gamma V_t^+(s_{t+1}) - V_t^+(s_t), & r_t \leq 0 \end{cases} \quad (7)$$

に基づき計算する。ここで、 γ は未来の報酬に対する割引率であり、 $V_t^+(s_t)$ は状態 s_t に対する積極的 critic モジュールの評価値である。同様に、警戒的 critic モジュールは、罰に基づく TD 誤差 δ_{t+1}^- を

$$\delta_{t+1}^- = \begin{cases} r_t + \gamma V_t^-(s_{t+1}) - V_t^-(s_t), & r_t < 0 \\ \gamma V_t^-(s_{t+1}) - V_t^-(s_t), & r_t \geq 0 \end{cases} \quad (8)$$

に基づき計算する。

$V_t^+(s_t)$ が正である場合、実行した行動 a_t は、報酬を最大化するという観点から望ましいものであったと考えられる。 $V_t^-(s_t)$ が正である場合、実行した行動 a_t は、罰を最小化するという観点から望ましいものであったと考えられる。これら二つが両立しうるかどうかは、環境に依存する。

それぞれの actor は、対応する critic から受け取った TD 誤差を元に、状態 s_t における行動 a_t の評価値を更新する。積極的 actor における評価値 $p_t^+(s_t, a_t)$ は、

$$p_{t+1}^+(s_t, a_t) = p_t^+(s_t, a_t) + \alpha \delta_{t+1}^+ \quad (9)$$

に従って更新され、同様に、警戒的 actor における評価値 $p_t^-(s_t, a_t)$ は、

$$p_{t+1}^-(s_t, a_t) = p_t^-(s_t, a_t) + \alpha \delta_{t+1}^- \quad (10)$$

として更新される。

また、critic モジュールは、各々の TD 誤差を基に、状態 s_t に対する評価値を

$$V_{t+1}^+(s_t) = V_t^+(s_t) + \beta \delta_t^+ \quad (11)$$

$$V_{t+1}^-(s_t) = V_t^-(s_t) + \beta \delta_t^- \quad (12)$$

に従って更新する。

エージェントのバイアスパラメータ e_t は、以下のよう
に確率的に修正することとする。まず、 $r_t \geq 0$ である
場合、バイアスパラメータは、確率 $k \cdot |r_t|$ で

$$e_{t+1} = e_t + \frac{1}{2}(1 - e_t) \quad (13)$$

として更新される。ここで、 k はバイアスパラメータの
変化確率を調整する定数であり、 $k > 0$ である。ただし、
 $k \cdot |r_t| > 1$ である場合には、確率 1 で式(13)を適用するも
のとする。

一方、 $r_t < 0$ である場合には、確率 $k \cdot |r_t|$ で

$$e_{t+1} = e_t - \frac{1}{2} e_t \quad (14)$$

として更新される。ここでも、 $k \cdot |r_t| > 1$ である場合に
は、 $r_t < 0$ の場合と同様、確率 1 で式(14)を適用するもの
とする。

なお、上記のいずれも成り立たない場合には、

$$e_{t+1} = e_t \quad (15)$$

である。

これにより、受け取る報酬の量が多ければ多いほど
バイアスパラメータが高くなり、積極的モジュールが
支配的になる。結果として、このような場合には積極
的な行動が促進されることとなる。一方で、受け取る
罰の量が多ければ、警戒的モジュールが支配的となり、
罰信号の回避が優先されることとなる。

なお、ここではバイアスパラメータ e_t は 0.5 を基準と
して 0 および 1 に向けて等しい割合で変動するものとし
たが、対象とする問題の性質によっては、式(13)および
式(14)右辺第二項の係数を変更する事によって、報酬と
罰のいずれが重視されやすくなるかを修正することも
可能である。ただし、そのような修正を行うには対象
となる問題についての予備知識が必須であることから、
本論文では汎用的な設定として報酬と罰とを対等に扱
うものとした。

4 計算機実験

4.1 実験環境

提案モデルの検証のために、以下のような実験を行
った。図 3 に示すような、グリッドによって区切られ
た平面環境をコンピュータ上で構築し、その中に強化

図 3 実験環境 A

Fig.3 Simulation environment A.

学習エージェント一体を配置した。エージェントは入
力として現在位置を正確に観測することができ、出力
として上下左右のいずれかに 1 グリッドの移動を行う
ものとした。

環境中には複数の餌が点在しており、また侵入時に
罰信号を受ける領域が存在する。エージェントがいず
れかの餌に到達した場合には、餌の大きさに応じた報
酬信号が与えられ、エージェントは初期位置に戻され
る。なお、エージェントが餌に到達しても、餌が失わ
れることはない。

エージェントが初期位置から動作を開始し餌に到達
するまでを 1 試行とし、全く学習がされていない状態
から、あらかじめ定めた回数の試行を完了するまでを
1 ランと呼ぶ。エージェントの学習は、1 ランを通じ
て常時行われる。

比較対象としては、2 で述べた従来型の actor-critic
モデル[1]を用いた。提案モデル、従来モデルとも、
actor の行動選択には前述の Boltzman 選択法を用い
た。

実験に用いたパラメータを表 1 に示す。バイアスパ
ラメータの変動係数 k を除いて、全てのパラメータは
提案モデルと従来モデルで同じ値を用いた。

表 1 実験パラメータ

Table 1 Simulation parameters.

Actor の温度パラメータ	T	1.00
Actor の学習率	α	0.10
Critic の学習率	β	0.10
TD 誤差の割引率	γ	0.90
バイアスパラメータの変動係数	k	0.10
バイアスパラメータ e_t の初期値		0.50

4.2 最善の方策の過程に罰信号が含まれる場合

まず、図3に示すような環境で行った実験について述べる。環境中には侵入時に罰信号を受ける領域が存在し、エージェントの初期位置から見て罰領域の手前に小さい餌、奥に大きい餌が配置されている。様々なパラメータを用いて比較実験を行ったが、そのいずれにおいても、罰領域を越えて大きい餌に到達した場合に、強化信号の総量が最大化されるように設定されている。

まず、この環境下において、強化信号の量の設定を変えて比較実験を行った。図4は提案モデルを用いた場合の結果、図5は従来モデルを用いた場合の結果である。どちらの場合にも、問題に特化した温度パラメータのスケジューリングは行わず、一回のランを通して

図4 提案モデルでの学習結果(強化信号の設定が異なる場合の比較)

Fig.4 The performance of the proposed model under different reinforcement settings.

図5 従来モデルでの学習結果(強化信号の設定が異なる場合の比較)

Fig.5 The performance of generic actor-critic model under different reinforcement settings.

て一定の温度パラメータを使用している。いずれも、連続する20回の試行ごとに最善の過程を選択した回数を集計し、それを100回のランについて平均したものをプロットした。

ここで、設定Aは、大きな餌の報酬を6.0、小さな餌の報酬を2.0、罰を-2.0に設定したものであり、同様に、設定Bはそれぞれを4.0/2.0/-1.0に、設定Cは6.0/2.0/-1.0に設定したものである。

従来モデルを用いた実験においては、小さな餌を獲得する次善の解のみが学習され、図5から判るように、大きい餌へと到達する最適な解が学習されることは稀であった。これは、罰領域を探索するリスクが極端に大きいのにに対し、小さい餌を通じて学習の初期段階から安定して報酬を獲得できてしまうためと考えられる。大きい餌から与えられる報酬が相対的に大きい場合(設定C)においては比較的良好な学習が行われたが、本質的には、この種の環境における従来モデルの性能は低いといえる。

一方、提案モデルを用いた実験においては、これら全ての設定において従来モデルに勝っている。また、従来モデルは成功率が強化信号の設定に極端に依存しているのに対し、提案モデルでは強化信号の設定への依存性が低く、安定して高い成功率を達成している。これは、提案モデルにおいては、積極性と警戒心の動的な調整により、小さい餌が安定して獲得できるようになると罰領域の探索が促進されるためと考えられる。

次に、従来モデルを用いて、温度パラメータのスケジューリングを行った場合の結果について述べる。試行回数を t 、冷却係数を C 、温度の初期値 T_0 とした時、試行回数 t の時点での温度 $T(t)$ を

$$T(t) = \begin{cases} T_0 - C \cdot t, & t \leq T_0/C \\ 0, & t > T_0/C \end{cases} \quad (16)$$

として、シミュレーテッドアニーリングを行った。大きな餌の報酬は6.0、小さな餌の報酬は2.0、罰は-1.0に設定した。初期温度 $T_0=100$ として、冷却係数 C を変えて比較した場合の結果を図6に、 $C=0.10$ として T_0 を変えて比較した場合の結果を図7に示す。

図より明かなように、適切なパラメータでアニーリングを行う事により、従来モデルでの最終的な学習精度は大幅に向上することがわかった。しかしながら、そのパラメータ設定は容易ではなく、また提案モデルと同等の学習精度を維持するためには十分にゆっくりと冷却を行う必要がある。

最後に、提案モデルを用いて、バイアスパラメータの変動係数を変えて比較実験を行った。図8に、この結果を示す。大きな餌の報酬は4.0、小さな餌の報酬は

図6 従来モデルでアニーリングを行った場合との比較(冷却係数が異なる場合の学習性能)

Fig.6 The comparison with generic actor-critic model with simulated annealing (varied cooling rate)

図7 従来モデルでアニーリングを行った場合との比較(初期温度が異なる場合の学習性能)

Fig.7 The comparison with generic actor-critic model with simulated annealing (varied initial temperature)

2.0、罰は-1.0に設定した。この図において、 $k=0.1$ の場合が図4の設定Bと同一の条件となっている。また、 $k=0.0$ の場合には、どのような報酬・罰が与えられてもバイアスパラメータ e_t が変動することはない、一方、 $k=1.0$ の場合には、いずれの報酬・罰が与えられた場合にも必ず e_t が変動することとなる。

この結果から、 e_t の変動係数 k が極端に低い場合には積極性の調整が充分に行われずに性能が悪化し $k=0$ において従来法とほぼ同等の性能となることが分かった。一方、それ以外の場合においては、変動係数に大きく依存することなく、安定した性能で学習を行うことができると言える。 e_t が常に変動するような極端な設定に

図8 提案モデルでの学習結果(バイアスパラメータの変動係数が異なる場合の比較)

Fig.8 The performance of the proposed model under different bias parameters.

においても、従来法以上の性能を示した。

4.3 より単純な環境における実験

本節では、より単純な環境で行った実験について述べる。二種類の環境について、それぞれ、提案モデルを用いて $k=0.10$ とした場合の結果と、従来モデルで $T=1$ として温度を固定した場合の結果、そして従来モデルで初期温度を100としてアニーリングを行った場合の結果の、三種類の結果を比較した。いずれも、連続する20回の試行ごとに最善の過程を選択した回数を集計し、それを100回のランについて平均したものをプロットした。

まず、図3に示した環境から、小さい餌を取り除いて行った実験について述べる。この設定下における結果を、図9に示す。これは、大きい餌の報酬は6.0に、罰は-1.0に設定し、アニーリングを行う場合の冷却係数は $C=0.20$ とした場合の結果である。この環境においては罰領域の手前に餌が存在しないため、従来モデルで温度を低く固定した場合には、最善解の学習が殆んど行われないという結果になった。一方で、従来モデルでアニーリングを行った場合と、提案モデルを用いた場合には学習が適切に行われた。

次に、最善の方策の過程に罰領域が存在しない環境下で行った実験について述べる。環境としては図10に示すものを用い、餌の報酬を6.0に、罰を-2.0に設定した。これにより、罰領域を迂回して餌に到達した場合に、強化信号の総量が最大化されることとなる。アニーリングを行う場合の冷却係数は $C=0.50$ とした。

この設定下における結果を、図11に示す。結果より明らかなように、このように学習が容易な環境においては、提案モデルと従来モデルはほぼ同等の学習精度

図9 提案モデルと従来モデルの比較(図3の環境から小さい餌を取り除いた場合)

Fig.9 The performances in an environment without the small food.

図10 実験環境 B

Fig.10 Simulation environment B

図11 提案モデルと従来モデルの比較(環境 B)

Fig.11 The performances in environment B.

を示した。このような環境では、提案モデルによる積極性と警戒心のバイアス調整は特に性能に影響を与えないと言える。

4.4 より状態数の多い環境における実験

本節では、図3の環境よりも状態数の多い環境として、図12に示す環境を用いた実験について述べる。二種類の餌と罰領域に加え、この環境には侵入不可能な壁が存在している。壁に向かって移動しようとした場合には、エージェントは単純に元の座標にとどまるものとする。大きな餌の報酬は6.0、小さな餌の報酬は2.0、罰は-1.0と設定した。

図13は、提案モデルでバイアスパラメータの変動係数を変えて比較を行った結果である。前述の実験と同様、連続する20回の試行ごとに最善の過程を選択した回数を集計し、それを100回のランについて平均したものをプロットした。図より明らかなように、この環境ではバイアスパラメータの変動がない場合には最善解の学習が全く不可能であり、積極性と警戒性の動的な調整による学習精度の向上が顕著に現れた。また、図14は、従来モデルでシミュレーテッドアニーリングを行い、冷却係数を変えて比較を行った場合の結果である。なお、この図における提案モデルの結果は、 $k=0.1$ の場合のものである。アニーリングを行わない場合に

図12 実験環境 C

Fig.12 Simulation environment C.

図13 提案モデルでの学習結果(環境 C)

Fig.13 The performance of the proposed model in environment C.

図14 従来モデルでアニーリングを行った場合の学習結果(環境 C)

Fig.14 The performances of generic actor-critic model with simulated annealing in environment C.

はやはり最善解の学習が不可能であり、また、図中では省略したが、冷却係数を0.1および0.05として実験を行った場合にも最善解の学習は全く行われなかった。冷却係数を0.01とした場合によりやうく提案モデルに近い学習精度が得られたが、図より明らかなように、学習の収束に必要な試行回数は非常に大きなものとなった。

5 おすび

本論文では、動物における積極性と警戒心に対応するような形で価値判断の変更を行う新しい強化学習モデルを提案した。このモデルは、積極的モジュールと警戒的モジュールに分けて学習を行い、二つのモジュールが行動選択に与える影響の比重を調整することで学習性能の向上を行うものである。

計算機実験により、最善の方策の過程に罰信号が含まれるような環境において、提案モデルが従来モデルに比べ安定して高い成功率で学習を行うことを確認した。これは、積極性と警戒心の調整が必要な環境において、提案モデルが有効に働いていることを示している。また、最善の方策の過程に罰信号が含まれない環境において、提案モデルが従来モデルと同等の性能を発揮することを確認した。従来モデルでシミュレートドアニーリングを使用した結果と比較した場合、最終的な精度では従来モデルにやや劣るものの、学習の収束に要する時間では提案モデルが大幅に優れているという結果が得られた。

今後の課題としては、より多くの問題設定における提案モデルの性能の検証と、複数モジュールを併用することによって上昇した計算コストの低減が挙げられ

る。また、提案モデルでは基本的に新しい経路を探索し続けるため、学習結果が完全には安定しないという特徴がある。これは、変化するような環境には適しているが、固定された環境では学習精度の悪化に繋がる。この点について、バイアスパラメータのスケジューリングによって改善が可能ではないかと考えている。

謝辞

本研究の一部は、文部科学省21世紀 COE プログラム「アクセス網高度化光・電子デバイス技術」の助成を受けて行われた。

参 考 文 献

- [1] R.S.Sutton and A.G.Barto, "Reinforcement Learning: An Introduction," MIT Press, 1998.
- [2] 清本盛明, 亀井且有, "部分観測マルコフ決定過程における位置ベクトルを用いた強化学習手法の提案," システム制御情報学会論文誌, Vol.14, pp.86-91, 2001
- [3] T.Jaakkola, S.P.Singh and M.I.Jordan, "Reinforcement learning algorithm for partially observable Markov decision process," Advance in Neural Information Processing Systems, 7, pp. 345-352, 1994.
- [4] 中島智晴, 有働昌代, 石渕久生, "ファジィ Q 学習によるサッカーエージェントの行動獲得," 日本知能情報ファジィ学会誌, vol.15, pp.702-707, 2003.
- [5] M.J.Mataric, "Reinforcement Learning in the Multi-Robot Domain," Autonomous Robots, vol. 4, pp.73-83, 1997.
- [6] O.L.Tinklepaugh, "An experimental study of representation factors in monkeys," Journal of Comparative Psychology, vol.8, pp.197-236, 1928.
- [7] L.P.Crespi, "Quantitative variation in incentive and performance in the white rat," American Journal of Psychology, vol.55, pp.467-517, 1942.
- [8] I.H.Witten, "An adaptive optimal controller for discrete-time Markov environments," Information and Control, vol.34, pp.286-295, 1977.
- [9] A.G.Barto, R.S.Sutton and C.W.Anderson, "Neuronlike adaptive elements that can solve difficult learning control problems," IEEE Transactions on Systems, Man, and Cybernetics, Vol.13, No.5, pp.834-846, 1983.
- [10] R.H.Crites and A.G.Barto, "An actor-critic algorithm that is equivalent Q-Learning," Advances in Neural Information Processing Systems, vol.7, pp.401-408, 1994.
- [11] C.J.Watkins and P.Dayan, "Q - Learning," Machine Learning, 8, pp.279-292, 1992.
- [12] J.Grefenstette, "Credit Assignment in rule discovery systems based on genetic algorithms," Machine Learning, vol.3, pp.225-245, 1988.
- [13] J.H.Holland, "Escaping brittleness: the possibilities of general-purpose learning algorithms applied to parallel rule-based systems," in

Machine Learning: An Artificial Intelligence Approach, vol. 2, Morgan Kaufmann Publishers, 1986.

(2004年5月5日 受付)

(2004年11月28日 採録)

[問い合わせ先]

〒223-8522 横浜市港北区日吉 3-14-1

慶應義塾大学 理工学部 情報工学科 萩原研究室

萩原 将文

電話：045-563-1151 (内線 43252)

ファックス：045-566-1747

E-mail: hagiwara@soft.ics.keio.ac.jp

著者紹介

たき た こういちろう
瀧田 航一朗 [非会員]

2000慶大・理工・情報卒。現在同大大学院博士課程在学中。ニューラルネットワークの研究に従事。

はぎわら まさふみ
萩原 将文 [正会員]

1982慶大・工・電気卒。1987同大大学院博士課程了。工博。同年同大助手。

現在、同大教授。ニューラルネットワーク、ファジーシステム、進化計算の研究に従事。1985年度丹羽記念賞、1986年度本会篠原記念学術奨励賞、1989年度IEEE Consumer Electronics Society 論文賞、1994年安藤記念学術奨励賞、1996年日本ファジィ学会著述賞受賞、IEEE、INNS、情報処理学会、電気学会各会員。

Reinforcement Learning Algorithm with Bias of Activeness and Cautiousness

by

Koichiro TAKITA and Masafumi HAGIWARA

Abstract :

In this paper, we propose a new reinforcement learning algorithm with a variable bias between two animalistic instincts - the activeness and the cautiousness. Although reinforcement learning methods are more flexible than supervised learning methods, it is difficult to determine the appropriate amount of reinforcement signals. For example, too much reward for the second best solutions can prevent the agent from exploring for the best solution, while insufficient amount of reward can prevent it from learning any kind of solutions. To overcome such problems, the proposed model uses two learning modules with a variable bias. One of the modules represents the activeness. Its sole goal is to maximize the amount of reward given from the environment. The other module represents the cautiousness. Its goal is to minimize the amount of penalty. By changing the bias between these two modules, the proposed model can perform efficient learning in wider variety of environments. With computer simulations, we confirmed that the proposed model can learn effectively in environments where conventional models show poor performance. In other environments, the proposed model showed performances comparable to conventional models.

Keywords : reinforcement learning, actor-critic, emotion

Contact Address : **Masafumi Hagiwara**

*Hagiwara Laboratory, Department of Information and Computer Science,
Faculty of Science and Technology, Keio University,*

3-14-1 Hiyoshi, Kohoku-ku, Yokohama-shi 223-8522, Japan

Tel: +81-(0)45-563-1151 (ext.43252)

Fax: +81-(0)45-566-1747

E-mail: hagiwara@soft.ics.keio.ac.jp