知能と情報（日本知能情報ファジィ学会誌）Vol.21, No.5, pp.722－733（2009）

原著論文

Shaping 強化学習を用いた自律エージェントの行動獲得支援手法†

前田 陽一郎 *1・花香 敏 *2

一般に，自律エージェントや自律移動ロボットに効率的な行動学習をさせるためには動物の学習メカニズムから工学的応用を行なうことは有効な手法であることが知られている．中でも，動物行動学，行動分析学や動物のトレーニング（調教）などで広く用いられている「Shaping」という概念が最近注目されている．Shapingは学習者が容易に実行できる行動から複雑な行動へと段階的，誘導的に強化信号を与え，次第に希望の行動系列を形成する概念である．本研究では繰り返し探索により自律的に目標行動を獲得できる強化学習にShapingの概念を取り入れたShaping強化学習を提案する．有効なShaping効果を検証するために強化学習の代表的なQ－Learning，Profit Sharing，Actor－Criticの3手法を用いた異なるShaping強化学習を提案し，グリッド探索問題のシミュレータを用いて比較実験を行なった．さらに，実際の動物などの調教の場などで知られている段階を追って行動を強化する「分化強化」という概念をShaping強化学習に取り入れた分化強化型ShapingQ－Learning（DR－SQL）を提案し，シミュレーション実験により手法の有効性が確認された．

キーワード：Shaping強化学習，分化強化，調教，自律エージェント，移動ロボット

1．緒　言

近年，ロボティックスの分野において，生物の認知プロセスを理解し，具体的なメカニズムとして具現化することの必要性が唱えられている[1]．例えば，認知ロボティクスでは作業の習熟過程や環境変化に呼応したロボットの経時的な変化を可能とするメカニズムが必要であると言われている．人間や動物は効率的にこれらを達成するために模倣や教示といった手法を用いている．これらをロボットに組み込んだ例として，石黒ら[2]は状態空間を効率よく構成するために人間から直接例が与えられる方法を提案している．また，野田ら[3]はやさしい状況から困難な状況に導くことで学習を加速する手法を提案している．

一般に自律エージェントや自律移動ロボットに効率的な行動学習をさせるためには生物の学習メカニズムから工学的模倣を行うことは有効な手法である．そこで，動物行動学，行動分析学や動物のトレーニングなどで広く用いられている“Shaping”という概念を自律エージェントの行動学習に応用することを考える．Shapingは学習者が容易に実行できる行動から複雑な行動へと段階的，誘導的に強化信号を与え，次第に希望の行動系列を形成する概念である．この概念はマウスを使った実験で数値的にその有効性が検証されている[4]．

Shapingの概念を工学的に応用した研究例として，M.Dorigoら[5]は拡張遺伝的アルゴリズムを用いた並列処理クラシファイアシステムの上段にShaping政策を用いた手法を提案している．椹木ら[6]は模倣学習したエージェントが常に理想的な行動によりタスクを達成できるようになる保証はないため，これを解決するためにエージェント自らが試行錯誤に基づく強化学習により修正していくというプロセスを一つの学習器上で実現する方法を提案している．A.Y.Ngら[7]は強化学習においてサブ報酬を与えると生じる局所ループの発生を回避するためにShaping報酬にポテンシャルの概念を取り入れた学習手法を提案している．いずれも興味深い研究であるが，Shapingの概念を学習システムに組み込んだ手法の体系的な検証がなされておらず，一般的な学習手法がまだ確立されていない．

本研究室ではこれまで，自律移動ロボットのための階層型ファジィ行動制御の研究を行っており，個々のサブタスクを記述したファジィルールを上位で切り替えるための行動選択ファジィルールを状態分割型強化学習を用いて自律的に戦略獲得する手法を提案してきた[8]-[10]．しかしながら，通常の強化学習のみでは

† Behavior Acquisition Supporting Method Used Shaping Reinforcement Learning for Autonomous Agent
Yoichiro MAEDA and Satoshi HANAKA

*1 福井大学大学院 工学研究科 知能システム工学専攻
Department of Human and Artificial Intelligent Systems, Graduate School of Engineering, University of Fukui
*2 村田機械株式会社 犬山R＆Dセンター京都分室
Reasearch & Development Division, Murata Machinery, Ltd.

良好な結果が得られるまでには膨大な試行を繰り返す必要があった．そこで，本研究では前述した Shaping の概念を強化学習に用いて，これまで考慮していなかった自律エージェントと人間との関わりなどの外界からのインタラクションの設計をすることにより自律エージェントの効率的な行動学習の実現を目指す．

本論文では様々な強化学習に Shaping の考えを取り入れた手法を提案し，自律エージェントの行動獲得に関するシミュレーション実験により性能評価を行なった．また，実際の動物などの調教の場では段階を追って行動を強化する「分化強化」という概念があり，効率良く目標行動を獲得させるには極めて有効であることが知られている．そこで，本研究ではさらにこの考えを取り入れた分化強化型 ShapingQ-Leaning（DR-SQL）を提案し，同様にシミュレーション実験によりその有効性を検証したのでこの結果についても併せて報告する．

2．行動分析学に基づく行動強化について

“Shaping”は，犬やイルカなどの動物の調教に広く一般的に使われている用語である．行動分析学でも行動を強化するための有効な概念として注目されている．Shaping とは行動を形成するという意味であるが，やらせたい行動に近い行動を強化しながら，少しずつ目標としている行動に近づけていくという概念である．

行動分析学では行動が強化されるときには必ず正の強化と弱化，負の強化と弱化が存在する．文献[4]では行動を強化する刺激や条件などで訓練者にとって愉快なものを「好子（報酬）」，不快なものを「嫌子（罰）」という用語を用いて，行動の強化を説明している．以下では好子の出現による Shaping について実例に基づいて説明する．

図 1 にイルカに芸を教えるときの Shaping について一例を挙げる．Shaping では初めに最終目標を決める．この例での最終目標はジャンプして輪をくぐるという行動である．最終目標を達成するために一連の行動を「分化強化」していく．分化強化とはいくつかに分割した行動を個々に強化していくことである．

まず，ジャンプすることを強化する．イルカの動きを観察して，たまたまジャンプしたときにエサを与える．エサが与えられることがイルカにとって好子（報酬）の出現になり，ジャンプする行動が強化される．ここで注意すべきことは好子を与えるタイミングと量である．強化したい行動の直後に好子を与えないとイルカはどの行動に対して好子が与えられたのかがわからなく，うまく強化されない．また，一回に与えるエサの量が多すぎても，イルカは行動が強化される前に満腹になってしまい，逆に少なすぎてもイルカは報酬

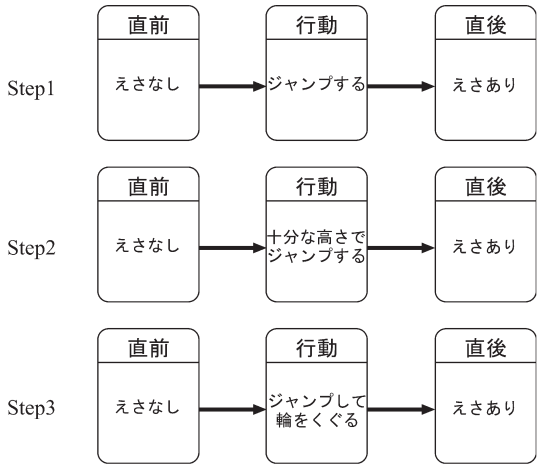


図 1　イルカにおける好子の出現による Shaping

に対して満足できなくて再度，同じ行動を取ろうと思わなくなる．イルカの訓練をはじめ，動物トレーナーとして長い経験を持つ生物学博士カレン・プライアも自身の書[11]で，調教の際には好子の提示，つまり強化は，変えようとする行動が起こったらすぐに与えなければならないと述べている．

次に，より高くジャンプしたときにエサを与える．これによりイルカは徐々に高くジャンプするようになる．最後にイルカの前に輪を用意し，輪をくぐった時にだけエサを与える．すると，イルカはジャンプして輪をくぐるという行動を学習することができる．このように徐々に目標行動に近い行動を分化強化していく方法が Shaping である．

また，Shaping を促進する方法として以下のようなものがあることが知られている．

(1) スキャニング（Scanning）

動物の行動を観察して，行なってもらいたい行動に少しでも近い行動が起きたら強化する方法である．この方法は偶然に犬が座ったときにえさをあげるとその行動が強化されて，“お座り”を覚えるといったように偶然によるため，複雑な行動を強化することは難しい．

(2) ターゲッティング（Targeting）

まず目印になるようなものを用意し，訓練者に目印を追従するように調教し，その目印を動かすことにより訓練者を誘導する方法である．この手法はアザラシなどの動物に芸を教えるときによく使われる．

(3) 身体的誘導法（Modeling）

学習させようとする行動をトレーナーの手を使って行わせることである．例えば，ゴルフのコーチが初心者のうしろから手を回してクラブを握り，それを振ってスイングさせるようなことが当てはまる．

(4)模倣(Mimicry)

他者の行動をまねることである．Shaping により新しい行動が獲得できた者の行動をうまく模倣したら強化する．イルカはお互いの行動を模倣する傾向が強いので，この手法が使われることが多い．

今回の提案手法では，これらの手法の中で人間が比較的容易に直接介入でき，かつ即効性がある「身体的誘導法」を導入することを考える．

3．Shaping 強化学習

従来の Q-Learning では環境との相互作用を通して，感覚入力と行動出力の両者のマッピングを記憶して学習を行ってきた．しかしながら，複雑な環境，複雑なタスクを対象とした場合，学習器が複雑かつ膨大になり，学習時間の増大，学習結果の再利用が困難であるなどの問題が生じる．そのため，これまでのエージェント内部で知覚した情報を効率的に学習する方法の設計(内部構造の学習)だけでは限界がある．そこで，本研究では前述した Shaping の概念を強化学習に取り入れ，自律エージェントと人間との関わりなどの外界からのインタラクションの設計(外部環境からの学習支援)を行なうことによりエージェントの効率的な行動学習を実現する Shaping 強化学習の提案を行う．

ここでは動物などの調教の場で広く使われている Shaping の概念を強化学習に取り入れた自律エージェントの効率的な行動獲得手法をいくつか提案する．従来の強化学習では最終目標が一定であるため，報酬関数等は固定して最初に定義されるのが一般的である．しかしながら，Shaping は学習者が容易に実行できる行動から複雑な行動へと段階的，誘導的に強化信号を与え，次第に希望の行動系列を形成する概念であるので，サブ目標が変化するため，これに基づいて自律エージェントに与える報酬関数等を人為的に変動させることができる．

今回用いる強化学習としては，代表的なQ-Learning，Profit Sharing，Actor-Critic の 3 手法を用いる[12, 13]．Shaping を加える方法は，エージェントの行動を人間が観察して，適宜，報酬関数等を人為的に変動させる方法と，条件を固定するためにあるサブゴールを達成すれば報酬関数を自動的に変動させる方法の 2 通りを用いた．以下にそれぞれについての提案手法を述べる．

尚，以下の手法では，人間がエージェントの行動を見て望ましい行動を取った時点で Shaping 報酬を与えることを考える．Shaping 報酬は人間が必要と思ったときにはいつでも一定報酬を与えられるものとする．

3.1 Shaping Q-Learning

Q-Learning は感覚入力(状態)と行動出力の組み合わせにおいて，各状態で可能な行動の中で将来にわたる行動評価値が最も高くなる行動をとるように学習を行う方法である．ここでは Shaping の要素を組み込む方法として 2 通りの提案をする．

(1)状態学習器に Shaping 報酬を付加する方法[SQL1]

図 2 (b)のように従来の Q 値の更新式を式(1)のように変形し，Shaping によるサブ報酬(Shaping 報酬 $S(t)$)を加える．Shaping 報酬とは従来のあらかじめ環境に設定されたサブ報酬とは異なり，人間が適宜，与えるサブ報酬と定義する．すなわち，Shaping 報酬を加えることにより人為的に Q 値を書き換えることができる．

$$\Delta Q(s_t, a_t) = \alpha(r_t + S(t) + \gamma \max_b Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)) \quad (1)$$

α：学習率　γ：割引率　r_t：報酬

$S(t)$：Shaping 報酬

if 人間が調教を行った時

$S(t) = c$(c は定数) $else\ S(t) = 0$

(2)行動選択器に Shaping 報酬を付加する方法[SQL2]

ここでは Q-table(Q値)と同様に状態と行動の関数である Shaping-table(S値)を用意し，行動選択の際にだけ参照する．この値は図 2 (c)のように人間がエージェントに調教を行なった時のみ値が更新され，Shaping の要素は直接 Q 値の学習(状態学習器)には反映されない．この値は学習の探索空間を絞り込む役割を担っている．ここでは式(2)のようにボルツマン選択に S 値を組み込む．S 値は Q 値とは異なり，逐次的に記憶されるマップであり，動物が調教により獲得した行動を考えて意識するのではなく，無意識で行動すること(体で覚える)と同様の効果が得られ，獲得行動が運動モデルとして確立される．

$$\pi(s_t, a_t) = \frac{\exp((Q(s_t, a_t) + S(s_t, a_t))/T)}{\sum_{b \in possibleactions} \exp((Q(s_t, b_t) + S(s_t, b_t))/T)} \quad (2)$$

$\pi(s_t, a_t)$：状態 s_t で行動 a_t を取る行動選択確率

T：温度定数(ボルツマン分布)

$S(s_t, a_t)$：Shaping-table(S 値)

if 人間が調教を行った時

$S(s_t, a_t) = S(s_t, a_t) + c$　(c は定数)

$else\ S(s_t, a_t) = S(s_t, a_t)$

3.2 Shaping Profit Sharing

Profit Sharing は感覚入力(状態)と行動出力のペアで記述されたルール系列に対して一括で報酬を与える

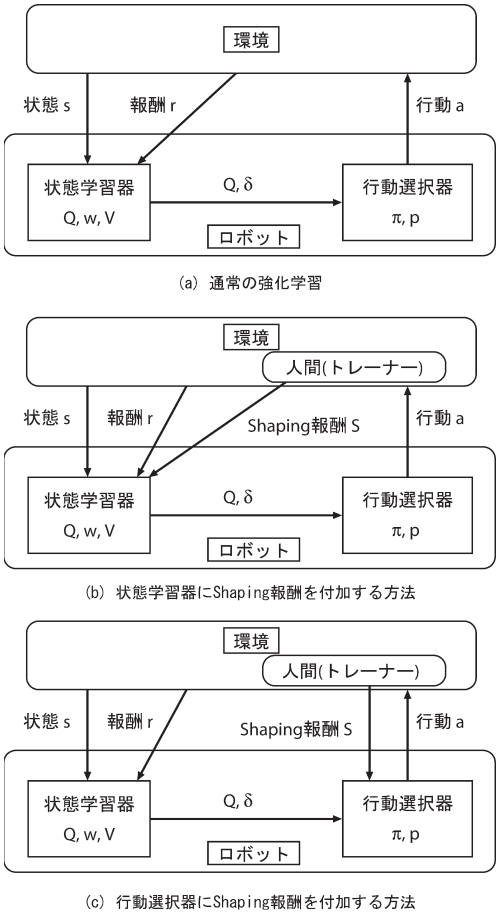


図 2　提案する Shaping 強化学習の概念図
（図中の記号の説明は本文中に記載）

ことにより連続した行動に対して効率的に学習を行なう手法である．報酬を獲得するまでのルール系列をエピソードと呼び，獲得報酬をエピソード内の各行動選択の評価値に分配し，エピソード単位で強化を行う．

(1)行動重みに Shaping 報酬を付加する方法[SPS]

Profit Sharing の報酬を分配する方法として，式(3)に示すような報酬を得た状態からどれだけ過去の状態であるかを示す h を引数とする強化関数 $f(h)$ を用いて，各ルールの重み w_{ik}（状態 $i=1,\ldots,l$；行動 $k=1,\ldots,m$）を式(4)に従って更新する．ここでは，Shaping 報酬は人為的に行動重み w_{ik} を変化させるのに用いられる．この手法はSQL1と同様の考え方であり，図 2 (b)に示す状態学習器に Shaping 報酬 $S(t)$ を与える手法である．Shaping 報酬が与えられた行動が重要なポイントと認識して，それまでに行なった行動にも意味があると考えて別途報酬を割り引いて与えることにより，連続したルール系列を効率よく学習する．

$$f(h) = \gamma^h (r_t + S(t)) \tag{3}$$

$$w_{ik} \leftarrow w_{ik} + \alpha f(h) \tag{4}$$

α：学習率　γ：割引率　r_t：報酬
$f(h)$：強化関数　w_{ik}：ルール重み

3.3 Shaping Actor-Critic

Actor-Critic は，行動空間が連続的な場合によく用いられる強化学習法で，状態価値を評価する Critic（状態評価部）と状態観測に応じて確率的に行動選択を行う Actor（行動決定部）が明示的に分かれている．Actor-Critic では，TD 誤差と呼ばれる見積もりと実際に行動した時に得られる評価値の誤差を用いて Critic の状態価値関数を更新し，さらにこの TD 誤差を用いて Actor の行動選択確率が更新される．

(1)Critic に Shaping 報酬を付加する方法[SAC1]

TD 誤差の出力に Shaping 報酬 $S(t)$ を加えることにより，式(5)に従ってTD 誤差を出力し，状態価値関数 $V(s_t)$ が更新される．この手法も SQL1や SPS と同様の考え方で，図 2 (b)に示す状態学習器にあたる Critic に Shaping 報酬が与えられる．エージェントは与えられた状態に重要な価値があると考える．

$$\delta_t = r_t + S(t) + \gamma V(s_{t+1}) - V(s_t) \tag{5}$$

δ_t：TD 誤差　γ：割引率　r_t：報酬

(2)Actor に Shaping 報酬を付加する方法[SAC2]

式(6)のように行動選択確率 $p(s_t, a_t)$ の更新式に Shaping 報酬 $S(t)$ を付加する方法である．Shaping 報酬を加えることにより人為的に行動価値関数が更新され，確率的探索に人間の意図を与えることができる．この式で，ステップサイズパラメータ β は大きいほど現時点の行動だけでなく近い過去にとった行動にも学習を反映させることができる．この手法は SQL2と同様の考え方で，図 2 (c)に示すように行動選択器にあたる Actor に Shaping 報酬が与えられている方法である．

$$p(s_t, a_t) \leftarrow p(s_t, a_t) + \beta(\delta_t + S(t)) \tag{6}$$

$p(s_t, a_t)$：状態 s_t で行動 a_t を取る行動選択確率
β：ステップサイズパラメータ

3.4 シミュレーション実験

これまでに提案した本手法の有効性を検証するためにグリッド探索シミュレータを作成し，シミュレーション実験を行なった．実験に用いた環境を図 3 に示す．Start から Goal へのグリッド上の経路探索を各強化学習法を用いて学習するが，エージェントに Shaping

を行なう場合，調教者(人間)がシミュレータ上のエージェントの動きを観察して，Joypad の十字ボタンを使って Shaping 報酬や Shaping-table の書き換えを適宜行なう．Joypad は調教者が Shaping 報酬を付与するタイミングで一定値を与えるために用いるが，Shaping-table の場合は調教者が行動(移動方向)を指示するために使用される．以下に本シミュレーションの条件を示す．

- 移動環境は10×10(初期環境)と20×20(複雑な環境)のグリッド空間(障害物あり，なし)とする．
- 自律エージェントのスタート(初期位置)は空間の左下，ゴールは右上とする．
- 自律エージェントの目標行動は最短経路でゴールへ到達することである．
- 自律エージェントは上下左右，斜めに進む行動をとることができ，合計 8 つの行動を各学習手法の行動選択規範に従って自律的に選択する．
- 自律エージェントはグリッド外には出ないものとし，外に出る行動を選択した場合は再度選択する．
- ゴールで得られる報酬を10×10のグリッドでは10，20×20のグリッドでは100とする．

実験 1（各種 Shaping 強化学習手法の比較）

3 章で提案した以下の 5 つの Shaping 強化学習手法の性能を比較するために図 3 (a)のような環境でシミュレーション実験を行なった．比較には通常の Q-Learning を用いた．Shaping 報酬は0.1を与えた．

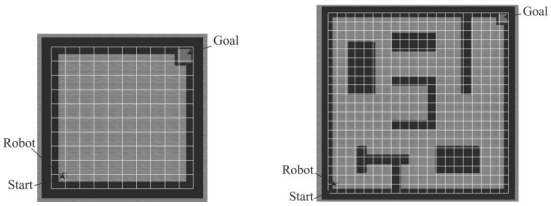
- Q-Learning の状態学習器＋Shaping 報酬(SQL1)
- Q-Learning の行動選択器＋Shaping 報酬(SQL2)
- Profit Sharing の行動重み＋Shaping 報酬(SPS)
- Actor-Critic の Critic＋Shaping 報酬(SAC1)
- Actor-Critic の Actor＋Shaping 報酬(SAC2)

実験 2（より複雑な環境での手法比較）

図 3 (b)のような20×20のグリッド(障害物あり)で学習手法間の比較実験を行なった．ここでは，実験 1 で比較的良い性能を示した Q-Learning と Actor-Critic に注目し，SPS を除く 4 つの Shaping 強化学習手法(SQL1, SQL2, SAC1, SAC2)と通常のQ-Learning を比較した．

実験 3（Shaping 報酬を記憶した場合の手法比較）

Shaping 報酬を毎回，人間が与えると一貫性がなく手法間の比較をする際に Shaping 報酬が不確定要素となるために，全く同じ条件で実験間の比較を行なうことは困難である．そこで，Shaping 報酬を与えた状態と行動を記憶しておきエージェントが自律的に同様の行動を行なったとき(特定のサブゴールで特定の行動を取ったとき)に，Shaping 報酬を自動的に与えた



(a)10×10：障害物なし　(b)20×20：障害物あり

図 3　実験に用いた環境

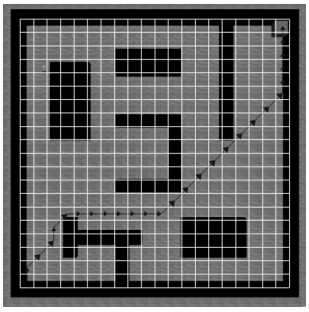


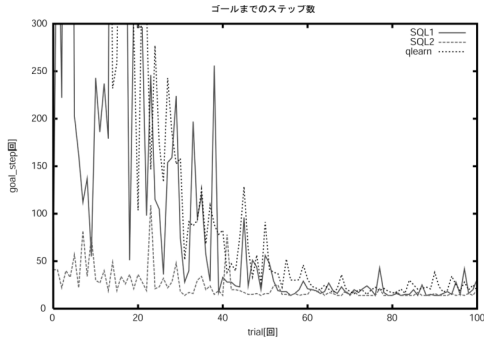
図 4　実験 3 で用いた Shaping 記憶マップ

場合の学習手法間の比較実験を行なった．実験には実験 2 で用いた 4 つの Shaping 強化学習と比較用に通常の Q-Learning を用いた．実験環境は実験 2 で用いた20×20のグリッド(障害物あり)で行なった．図 4 に調教者が事前実験で Shaping 報酬を与えた場所(状態)と方向(行動)をマップとして記憶した Shaping 記憶マップを示す．今回はエージェントが最短経路を進む行動が獲得できるように Shaping 記憶マップを作成した．エージェントが記憶したマップ上の移動方向に行動をしたときにのみ，一定の Shaping 報酬が状態学習器，行動選択器のそれぞれに付加される．

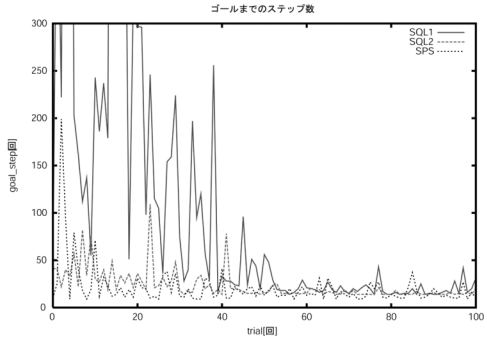
3.5　実験結果

実験 1 ～ 3 の結果を図 5 ～ 7 に示す．グラフの横軸は試行回数，縦軸はゴール到達までのステップ数を示す．

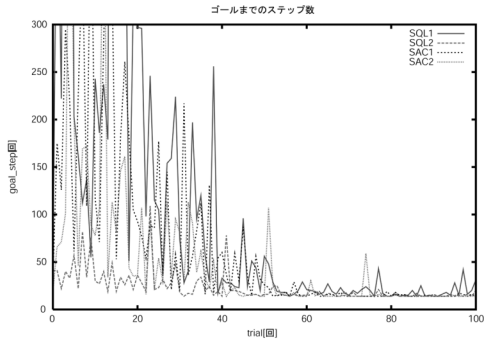
実験 1 より 5 つの Shaping 強化学習の中で比較的性能が良かったのは行動選択器に Shaping 報酬を付加する方法(SQL2)と Actor に Shaping 報酬を付加する方法(SAC2)であった．これらの手法は，行動に対して Shaping を与えている点で，共通しており，調教では強化信号を与えるタイミングは学習者がその行動を行った直後にどの行動に対して強化が与えられたかを明確にすることが重要であることが知られていることに対応している[4]．これらの手法は行動分析学，トレーニング法の観点から見ても適した方法であると考えられる．



(a) SQL



(b) SPS



(c) SAC

図 5 実験 1 の結果

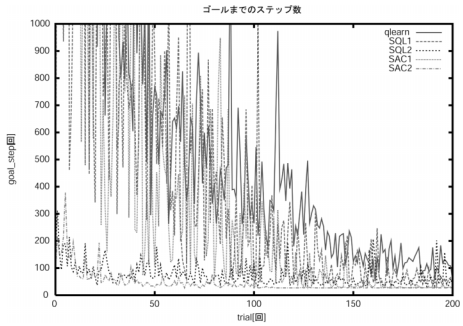


図 6 実験 2 の結果

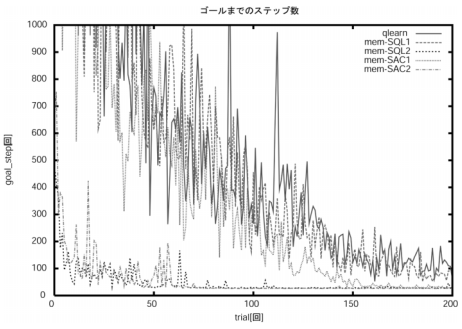


図 7 実験 3 の結果

逆に，状態学習器に Shaping 報酬を付加する方法であるSQL1，SAC1は Shaping 報酬を至る所で与えてしまうと局所解に陥りやすいことがわかった．Shaping 報酬を任意に与えているので最適性の保障も崩れ，Shaping 報酬の量，与え方によっては準最適解で収束したり，結果が振動的になったりする様子が見られた．尚，SPS については，SQL2やSAC2とほぼ同等の収束性能を示したが，収束後の結果が安定せず，SQL1やSAC1と同じく振動的な結果となった．

実験 2 のより複雑な環境下の場合においても，実験 1 と同様にSQL2，SAC2が良好な学習性能を示した．しかし，複雑な環境になると Shaping 報酬を与える回数が増え，調教者の負担が大きくなる．そのため，行動政策を持たないエージェントを初めから調教をするのではなく，ある行動政策を持ったエージェントの行動政策修正に人間を介在させるなどの手法を考える必要がある．

実験 3 の結果は図 7 であるが，図中の mem がついたものは Shaping 記憶マップを用いた実験結果を示す．この実験でもこれまでの結果と同様にSQL2，SAC2が良い結果を示した．これにより同じ様な Shaping 報酬は行動選択器に与えた方が良い結果を得ることがわかった．あらかじめ獲得したい行動の道筋がわかっている場合，これを事前に Shaping 報酬として与えておけば，人間が介在しないため，学習の自動化が可能である．しかし，今回のような最短経路を進む行動が獲得できるような Shaping を記憶したマップでは Shaping が与えられていない地点が多く存在して，エージェントがその地点に達したとき通常の Q-Learning を行なっているのと同様にランダム探索をするので大きな時間を有する試行が存在する．これを解消するために多くの点で Shaping を記憶しようとすると逆に上述のような調教者への負担が大きくなるという問題が存在する．

4．分化強化型 Shaping 強化学習

前章で提案した Shaping 強化学習は目標行動までの学習の途中で Shaping 報酬を与え，自律エージェントを誘導していく手法であった．しかしながら，この手法では学習初期から目標行動を目指して学習を行っていくので問題が複雑になるほど学習効率が悪くなる．これは，何もできないイルカにいきなり高い位置に輪を見せてくぐれと言っているのと同等である．3章で紹介したように初めから目標行動を目指すのではなく，まずはジャンプすることから覚えさせるといったように徐々に行動を強化していく「分化強化」（Differential Reinforcement）という概念を基に調教を行なっていくほうが効率的な学習が行なえる．そこで本章では，前章で述べた Q-Learning を用いた Shaping 強化学習で得られた知見を基に，複雑な環境下でも段階的に効率良く学習のできる分化強化型 Shaping Q-Learning（Differential Reinforcement-type Shaping Q-Learning: DR-SQL）を提案する．

4.1 分化強化型 ShapingQ-Learning（DR-SQL）

自律エージェントの行動学習に分化強化の概念を組み込むにはいくつかの方法が考えられるが，本研究では目標行動を人間の手によりいくつかの行動に分割して初期状態に近い行動から徐々に行動学習を行なっていく手法を考える．分割した行動にはそれぞれのサブ目標行動（サブゴール）があり，その行動が達成されたときに「サブゴール到達報酬」を与える．ここでのサブゴール到達報酬とは前章で述べた Shaping 報酬とは違い，環境のある場所（サブゴール）に固定であらかじめ設定された報酬であり，これを用いてサブゴールまでの行動学習を通常の強化学習と同様に行なう．

自律エージェントはランダム探索により行動学習を行ない，サブゴールに到達したらサブゴール到達報酬を与え，初期位置に戻して分割した行動を繰り返して，学習を行なう．サブ目標行動が獲得されたら，次に新たなサブ目標行動となる行動学習に移行する．このような行動学習を行なう際に，前章の Q-Learning を用いた Shaping 強化学習で得られた知見を基に，DR-SQL の学習システムを構築する．

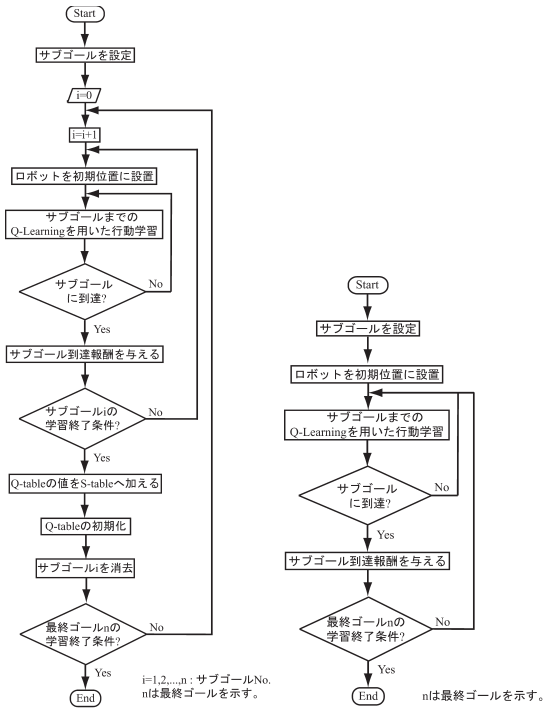
Shaping 強化学習では，サブゴール到達報酬を与えて自律的にエージェントに行動獲得させると Shaping 報酬を至る所で与えてしまうと局所解に陥りやすいという問題が生じることがわかった．そこで，サブゴールまでの行動学習が行なわれたら，それまで学習した結果を別のメモリに格納することを考える．例えば，Q-Learning では S-table というもの定義し，学習結果を格納する．S-table とは SQL2で定義した Shaping-table と同様の概念のもので，Shaping により学習された値を格納するものである．S-table は状態と行動の関数で Q-table と同じ型を持ち，Shaping-table と同様に行動選択の際に参照され，行動に反映される．

この手法の比較手法として，DR-SQL とは異なり分化強化ではなく，サブゴールに到達したら単純にサブゴール到達報酬のみを与える Q-Learning（以後，Sub-QL と呼ぶ）を考える．この手法は，サブゴールに達してもエージェントを初期位置に戻さずに学習を続け，サブゴールの設定場所は DR-SQL と同様である．

4.2 DR-SQLの学習アルゴリズム

前述の Q-Learning を用いた分化強化型 Shaping 強化学習の考え方を基に DR-SQL のアルゴリズムを構築した．その学習手順を以下に，アルゴリズムフローを図 8 (a) に示す．また，比較のために用いた Sub-QL のアルゴリズムフローを図 8 (b) に示す．

Step 1 目標行動までのタスクにいくつかのサブゴールを設定する．これは人間（調教者）があらかじめ適切に決める．（i＝0）



(a) DR-SQL　　(a) DR-SQL

図 8　実験におけるアルゴリズムフロー

Step 2 エージェントはサブゴールに到達したらサブゴール到達報酬が与えられ，サブゴールまでの学習を Q-Learning を用いて行なう．

Step 3 サブゴール i までの学習が収束してきたら学習を終了する．終了条件は人間が判断する方法と機械的に判断する方法が考えられるが，次節の実験にて比較する．

Step 4 サブゴール i までの学習が終了したら，学習した Q-table を S-table に加算する．

Step 5 Q-table を初期化し，サブゴール i を消去する．

Step 6 学習終了条件(ゴール＝サブゴールn)を満たせば，学習を終了する．そうでなければ，次のサブゴール(i+1)への学習に移行し，**Step 2** へ戻る．

この手順で Shaping 強化学習を行なうとエージェントは段階的に効率良く目標行動に近い行動を取るようになり，最終的には比較的短い学習時間で目標行動が獲得される．

4.3 シミュレーション実験

分化強化型 Shaping 強化学習(DR-SQL)の有効性を検証するために，前章の実験で用いた図 3 (b)と同じ20×20のグリッド(障害物あり)探索問題を学習した．シミュレーションに用いたエージェント，環境等の条件は前章の Shaping 強化学習のときと同等である．以下では本シミュレーション実験で行なった各実験条件について述べる．

実験 4 (Q-Learningとの比較実験)

前述のDR-SQLアルゴリズムに従い，図 9 のような 4 つのサブゴールを設定した．これはゴールまでのいくつかあるルートの内，障害物により作られた分岐点の入り口になるように設定した．比較には通常の Q-Learning と Sub-QL を用いた．

DR-SQL においてサブゴールまでの学習の終了判定は，エージェントが i 番目のサブゴール学習試行時にサブゴール到達までに要したステップ数(総移動距離)を $s(i)$ として，式(7)のように過去 5 回の $s(i)$ の増減の差の平均が 1 未満になった場合とした．これはほぼ同じ経路を繰り返したどるようになったときを表している．実験 5 ではこの終了判定に人間を関与させたが，人間が終了判定を下す場合，すべての実験において同じ条件にすることが不可能なため，条件を統一した比較実験を行なう目的で，ここでは自動終了判定を用いた．

$$\frac{\sum_{i=1}^{5} |s(i) - s(i-1)|}{5} < 1 \qquad (7)$$

実験では各パラメータの初期値は，学習率を0.3，割引率を0.8，温度定数を0.3，報酬を100，DR-SQL のサブゴール到達報酬を 5 ，Sub-QL のサブゴール到達報酬を0.1に設定した．

実験 5 (異なるサブゴール終了判定の比較実験)

サブゴールの終了判定を人間が自律エージェントの動きを観察して，良くなったと判断したときに終了する方法(以後，human-DR-SQL と呼ぶ)と前述の実験 4 の終了判定により自動終了する方法との比較を行なう．条件を統一した比較実験を行なうため実験 4 では自動終了を行なったが，本来の動物の調教で用いられている分化強化ではトレーナーの判断により次の段階へと進んでいく．この判断には調教者達の間では一定のルールはあるものの絶対的なルールは存在しない．その都度変動していき，優秀な調教者ほど直感的に好判断を下すことができる．

実験 6 (サブゴール，サブ報酬の設定差による比較実験)

ここでは，サブゴール到達報酬の量，サブゴール数，サブゴールの設定場所の違いによる性能の評価実験を行なう．実施した実験の条件を以下の表に，サブゴールの設定場所を図10に示す．

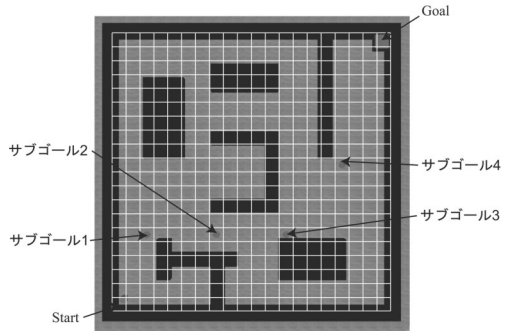
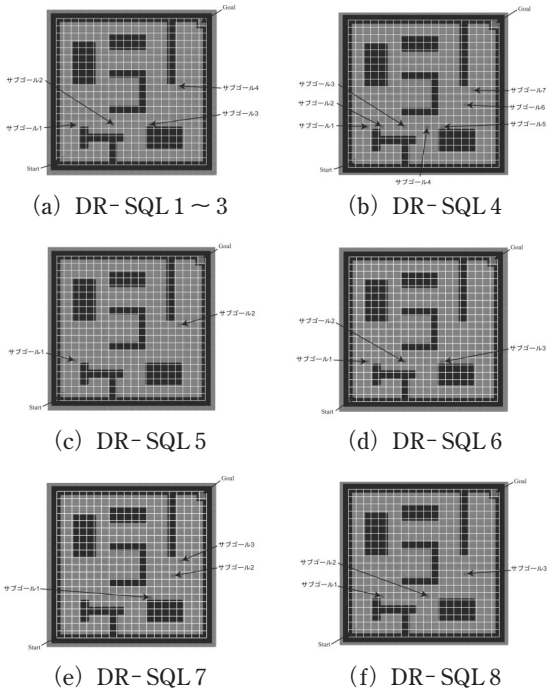


図 9 実験 4 でのサブゴール設定

表 1 実験 6 の実験条件（SG:サブゴール）

実験	SG 到達報酬の量	SG 数	SG の設定場所
DR-SQL1	5.0	4	図 10(a)
DR-SQL2	10.0	4	図 10(a)
DR-SQL3	50.0	4	図 10(a)
DR-SQL4	5.0	7	図 10(b)
DR-SQL5	5.0	2	図 10(c)
DR-SQL6	5.0	3	図 10(d)
DR-SQL7	5.0	3	図 10(e)
DR-SQL8	5.0	3	図 10(f)



（a）DR-SQL 1 ～ 3　（b）DR-SQL 4

（c）DR-SQL 5　（d）DR-SQL 6

（e）DR-SQL 7　（f）DR-SQL 8

図 10　実験 6 でのサブゴール設定

4.4 実験結果

図11～13に DR-SQL の実験結果を示す．グラフの横軸は試行回数，縦軸はゴール到達までのステップ数を示す．

実験 4 の結果を図11に示す．DR-SQL はサブゴールへ到達した場合に初期位置に戻るので，試行を 1 回と数える．試行の 1 回目から図中の横軸上の矢印の時点までが最後のサブゴール（最終ゴールの 1 つ前のサブゴール）へ到達したときの試行回数を表し，矢印以降が最終目標であるゴールを目指していることを示す（以後の DR-SQL の結果には同様の矢印を記す）．結果は Q-Learning が500回試行を終えても収束していないが，DR-SQL は200回試行以内でゴールできる状態に収束している．また，ゴールへ500回到達したときの総ステップ数は，Q-Learning が平均150210ステップで，Sub-QL が平均131973ステップ，DR-SQL が30883ステップであった．DR-SQL が1/4以下の短い時間で収束しており，良い性能を示している．

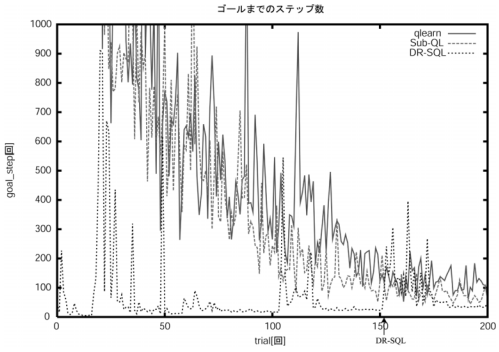


図 11　実験 4 の結果

実験 5 の結果を図12に示す．人間がサブゴールの終了判定を判断している human-DR-SQL は途中のサブゴールの学習に多少時間がかかっているが，最終的な

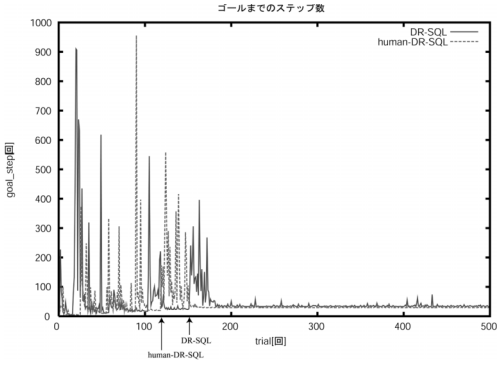
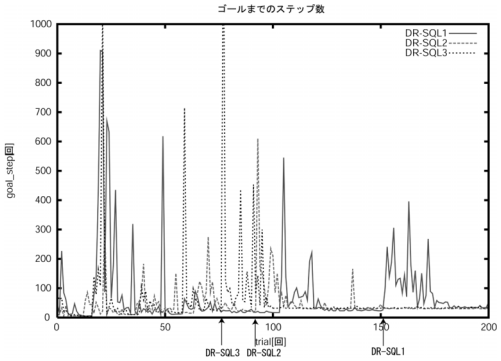
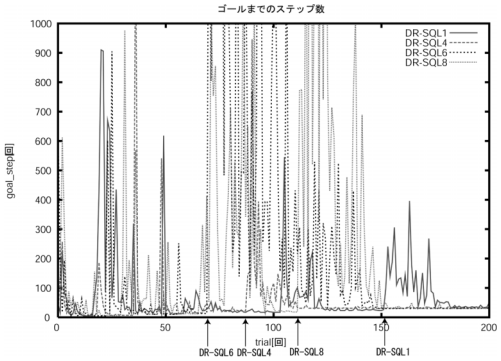


図 12　実験 5 の結果



（a）サブ報酬の大きさによる比較



（b）サブ報酬の設定場所、数による比較

図 13　実験 6 の結果

表 2 実験 6 の結果（総ステップ数）

実験	総ステップ数 (回)
Q-Learning	131973
DR-SQL1	30883
DR-SQL2	21823
DR-SQL3	24030
DR-SQL4	22496
DR-SQL5	ゴール未到達 (サブゴール 2 で局所解)
DR-SQL6	59319
DR-SQL7	ゴール未到達 (サブゴール 1 で局所解)
DR-SQL8	58424

ゴールの学習は終了判定を自動的に判定している方法よりも早い時間で収束していることがわかった．

実験 6 の結果を図13に示す．図13(a)にはDR-SQL1からDR-SQL3までの結果を示し，図13(b)にはDR-SQL1とDR-SQL4，DR-SQL6，DR-SQL8 の結果を示す．DR-SQL5はサブゴール 2 まで，DR-SQL7 はサブゴール 1 までの学習で局所解に陥ってしまい，最終ゴールにまで達することなく目標行動を獲得することができなかったため，図13(b)のグラフには結果を記載しなかった．DR-SQL1〜3の結果により，サブゴール到達報酬の量が多いほど最後のサブゴール到達までの学習時間が早いが，場合によっては局所解に陥ってしまい性能が悪くなることがわかった．サブゴールの設定場所，数による違いは最も細かくサブゴールを設置したDR-SQL4が表 2 に示すように他のDR-SQL1と比べても最も良い性能を示した．

5．考察

実験 4 により従来の Q-Learning，Sub-QL より分化強化型 Shaping Q-Learning(DR-SQL)の方が格段に良い性能を示すことがわかった．始めのサブゴール地点を指定するだけであるので調教者の負担は前章の Shaping 強化学習手法に比べると非常に軽く，性能も良いことがわかった．エージェントがサブゴールに達した経路に対しては関与していないため，ゴールから戻るような行動にも報酬が与えられることがたまに生じた．確率的にはスタート地点から近い方からサブゴールに達するので今回は問題がなかったが，本来の調教ではサブゴールに達したときはいつでもサブゴール到達報酬を与えるわけではない．時には罰を与えることもあるが，本論文では議論していないので今後，検討する必要があると考える．

実験 5 では，人間が終了判定を行なう方が良いことが示された．今回，終了判定を自動判別する条件はかなり厳しく，過去 5 回のステップ数の偏差平均が1未満で何度もサブゴールに到達するようにならないと終了しないように設定したが，人間が判断する場合，サブ目標行動が獲得できたときに学習を終了する傾向が強かった．また，人間が判断を行なう場合，すべてのサブゴールで同じような判別をするわけではなく，スタート地点から遠ざかるほど，判定条件が緩む傾向がある．これは，スタート地点から遠ざかるほどサブ目標行動の獲得に時間がかかり，ある程度の精度で妥協する傾向がある．ゴール地点に近いサブ目標行動は目標行動の学習により得られる報酬の伝播をすぐに受けることができるのでサブ目標行動を終了してもよいと考えられる．このような人間が調教を行なった場合のヒューリステックスを取り込むことで，DR-SQL のさらなる性能向上が期待できる．

サブゴール到達報酬の量を大きくしてもDR-SQL1〜3では最終的な結果にはそれほど顕著な差が見られなかった．サブ目標を設定したことにより，エージェントが一回に学習する空間が狭くなったので最低限の報酬量を持っていれば十分にサブ目標行動を獲得することができる．学習時間は総合的にはそれほど変わらず，最も余計な時間がかかるのはサブ目標から大きく離れた行動を取ったときである．実際の調教でも全く違った行動を取ったときは罰を与えたり，学習をやめさせ，元の位置に戻したりする．このような処理をこの手法に取り入れると，より短時間で学習ができるのではないかと考えられる．

実験 6 の DR-SQL5，DR-SQL7のようにサブゴール間の距離が長くなりすぎると，その間の学習に時間がかかったり，サブゴール到達報酬の量が少なく十分に伝播されないことがわかった．特にサブゴール間に分岐などの意思決定を必要とするような環境が存在するとこのような傾向になるので，分岐点などのポイントに必ずサブゴールを設定することが重要である．また，サブゴールの数が多くなると，短時間でより調教者の意図する行動に近い行動を取ることがわかった．サブゴールの数が多くなると新しく学習する学習空間が狭まるので 1 つのサブ目標行動の学習にかかる時間も短くて済むことがわかった．今回はサブ目標行動の終了判定を自動的に行なったが，人間が判断する場合はサブゴールが増えると調教者の負担がかかる．実際の調教も始めはサブゴールの数は少なく取り，調教者の負担を軽くしようとする．学習がうまくいかないときはサブゴールの数を増やし，細かく段階を追って調教をしていく．自律エージェントの調教でもサブゴールの数，設置場所，量などを可変で行なえるようにすれば，より複雑な環境下でも柔軟に対応できるのではないかと考えられる．

前章の Shaping 強化学習の実験では，エージェントが望ましい行動を取った時に人間が Shaping 報酬を

その都度与えているため，今回の DR-SQL の実験と比べるとかなり多くの報酬がエージェントに与えられており，当然最終ゴールへ到達する時間も早く（goal_step 数も少なく）なっている．ただ，Shaping 強化学習の場合，複雑な環境になると Shaping 報酬を与える回数が増え，調教者の負担が増大するという問題があった．これを改善するために，サブゴールに到達するたびに自動的にサブゴール到達報酬を与える DR-SQL の実験を行ったが，この手法では，人間が与えるのは最初のサブゴールの場所のみで報酬付与や終了判定（実験 5 のみ人間が判定）は自動で行っている．そのため，報酬がサブゴールに到達した限られた時のみ与えられるので，前章の Shaping 強化学習の実験に比べるとトータルな報酬量は少なく，当然最終的なゴール到達性能は下がる．しかしながら，人間への負荷という点では圧倒的に改善されており，いずれの手法を選ぶかはシステム設計者の判断に任されている．

6．結言

本研究では Shaping 強化学習を用いた自律エージェントの行動獲得に有効な手法を提案した．動物のトレーニングなどで広く用いられている Shaping をエージェントの行動学習に用いることが有効であることがわかった．さらに，複雑な環境下でも実際の動物などの調教の場で使われている段階を追って行動を強化する「分化強化」の概念を Shaping 強化学習に取り入れた分化強化型 Q-Learing を提案することによって調教者（人間）の負担を増やさずに効率的な行動学習ができることもわかった．

今回はグリッド探索問題だけの検証にとどまったが，本手法は他の複雑な環境における行動獲得問題にも応用でき，Shaping を用いることでインタラクティブに効率的な学習が可能となる．ロボットや自律エージェントが人間と共存していく中で，コミュニケーションを取りながら様々なタスクをこなすことができる一つの有効な手法となるのではないかと考える．また本手法では，Shaping 報酬の大きさ，与えるタイミングなどの報酬関数の設計やサブゴールの分割方法，サブ目標行動の終了条件など考慮すべき問題が多く存在する．これらは実際の調教の場でも重要な鍵を握っており，行動分析学や動物のトレーニングなどでは現場におけるいくつかのノウハウとしての解決方法が存在している．これらを参考に，より複雑な環境やタスクでも効率的な学習が可能なシステムを目指していく必要があると考える．

参 考 文 献

[1] 浅田稔，石黒浩，國吉康夫，“認知ロボティクスの目指すもの，”日本ロボット学会誌，Vol.17，No.1，pp.1-5（1999）

[2] H.Ishiguro, R.Sato, T.Ishida,” Robot oriented state space construction,” Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems（IROS’96），pp.1496-1501（1996）

[3] 野田彰一，浅田稔，俵積田健，細田耕，”強化学習によるロボットの行動獲得の効率化に関する考察－簡単なタスクからの学習LEM－，”第 4 回ロボットシンポジウム予稿集，pp.67-72（1994）

[4] 杉山尚子，島宗理，佐藤方哉，リチャード・W・マロット，マリア・E・マロット，行動分析学入門，産業図書（1998）

[5] M.Dorigo, M.Colombetti,” Robot shaping: Developing autonomous agents through learning,” Artificial Intelligence 71, pp.321-370（1994）

[6] 田淵一真，谷口忠大，椹木哲夫，”模倣学習と強化学習の調和による効率的行動獲得，”The 20th Annual Conference of the Japanese Society for Artificial Intelligence, pp.212-215（2006）

[7] A.Y.Ng, D.Harada, S.Russell,” Policy invariance under reward transformations: Theory and application to reward shaping,” Proc. of the Sixteenth ICML. Morgan Kaufmann（1999）

[8] Y.Maeda, S.Hanaka, W.Shimizuhira,” Multi-Layered Fuzzy Behavior Control Method for Autonomous Soccer Robot with MOVIS,” Proc. of the 3rd International Symposium on Autonomous Minirobots for Research and Edutainment（AMiRE 2005），pp.125-132（2005）

[9] 花香敏，前田陽一郎，”Shaping 強化学習を用いた自律移動ロボットの行動獲得，”第18回ファジィ・コンピューティング研究部会ワークショップ，05-18（2006）

[10] Y.Maeda and W.Shimizuhira,” Multi-Layered Fuzzy Behavior Control for Autonomous Mobile Robot with Multiple Omnidirectional Vision System: MOVIS,” Journal of Advanced Computational Intelligence and Intelligent Informatics（JACIII），Vol.11, No.1, pp.21-27（2007）

[11] カレン・プライア，うまくやるための強化の原理，二瓶社（1998）

[12] R.S.Sutton and A.G.Barto, Reinforcement Learning: An Introduction, The MIT Press（1998），三上貞芳，皆川雅章共訳，強化学習，森北出版（2000）

[13] 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，人工知能学会誌，Vol.14，No.5，pp.800-807（1999）

（2009年 3 月 1 日　受付）
（2009年 8 月 7 日　採録）

[問い合わせ先]
〒910-8507　福井県福井市文京3-9-1
福井大学大学院 工学研究科 知能システム工学専攻
前田　陽一郎
TEL：0776-27-8050
FAX：0776-27-8050
E-mail：maeda@ir.his.u-fukui.ac.jp

著 者 紹 介

前田 陽一郎［正会員］

1981年大阪大学基礎工学部機械工学科卒業．1983年同大学院基礎工学研究科修士課程修了．同年，三菱電機(株)入社．中央研究所，応用機器研究所，産業システム研究所を経て，1989年から1992年まで通産省技術研究組合国際ファジィ工学研究所(LIFE)へ出向．1995年より大阪電気通信大学工学部経営工学科を経て，総合情報学部情報工学科助教授．博士(工学)．1999年から2000年までカナダ・ブリティッシュコロンビア大学(UBC)客員研究員．2002年福井大学工学部知能システム工学科助教授，2007年同大学大学院工学研究科知能システム工学専攻教授，現在に至る．主として，ソフトコンピューティング手法による自律ロボットの知能化研究，および人とロボットの双方向インタラクションに関する人間共生システム研究に従事．計測自動制御学会，日本ロボット学会，人工知能学会，日本感性工学会などの会員．

花香 敏［非会員］

2005年福井大学工学部知能システム工学科卒業．2007年同大学大学院工学研究科知能システム工学専攻博士前期課程修了．同年，村田機械株式会社へ入社．現在に至る．

Behavior Acquisition Supporting Method Used Shaping Reinforcement Learning for Autonomous Agent

by

Yoichiro MAEDA and Satoshi HANAKA

Abstract：

Generally, it is known that the engineering application simulated from the learning mechanism of animals is useful to make learn behaviors of the autonomous agents or mobile robots efficiently. Above all, a general idea of "shaping" used by ethology, behavior analysis or animal training is a remarkable method recently. "Shaping" is a general idea that the learner is given a reinforcement signal step by step gradually and inductively forward the behavior from easy tasks to complicated tasks. In this research, we propose a shaping reinforcement learning method took in a general idea of "shaping" to the reinforcement learning that can acquire a desired behavior by the repeated search autonomously. Three different shaping reinforcement learning methods used Q-Learning, Profit Sharing, and Actor-Critic to check the efficiency of the shaping were proposed and the experiment by the simulator of grid search was performed. Furthermore, we proposed the Differential Reinforcement-type Shaping Q-Learning (DR-SQL) applied a general idea of "differential reinforcement" to reinforce a special behavior step by step such as real animal training, and confirmed the effectiveness of this method by the simulation experiment.

Keywords：Shaping Reinforcement Learning, Differential Reinforcement, Amimal Training, Autonomous Agent, Mobile Robot

Contact Address：**Yoichiro MAEDA**

Department of Human and Artificial Intelligent Systems, Graduate School of Engineering, University of Fukui

3-9-1, Bunkyo, Fukui-shi, Fukui 910-8507, JAPAN

TEL & FAX：0776-27-8050

E-mail ：maeda@ir.his.u-fukui.ac.jp