A3-1 Shaping 強化学習を用いた自律移動ロボットの行動獲得に関する研究

福井大学大学院 工学研究科　知能システム工学専攻
花香　敏　(指導教官：前田 陽一郎 助教授)

1　緒 言

近年、ロボット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することを目指す認知ロボティクスの必要性が唱えられている [1]。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 [2]。また、浅田ら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 [3]。これは環境設定自体を提供することに対応しており、学習結果を外部観測によって評価することで状況の切り替えを制御している。

一般に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 という概念をロボッ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]。

本研究室ではこれまで、階層型ファジィ行動制御 [5] において、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。しかしながら、通常の強化学習のみでは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る。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、これまで考慮してなかったロボッ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ロボットの効率的な行動学習を実現する [6]。本論文では様々な強化学習の手法に Shaping の考えを取り入れて自律移動ロボッ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ので、その結果について報告する。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、本研究ではさらに代表的な強化学習法である Q-Learning にこの考えを取り入れた分化強化型 Shaping Q-Leaning(Differential Reinforcement-type Shaping Q-Learning: 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合わせて報告する。

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、やらせたい行動に少しでも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、弱化、負の強化、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」、不快なものを「嫌子」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 にはイルカに芸を教えるときの Shaping について例を挙げる。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

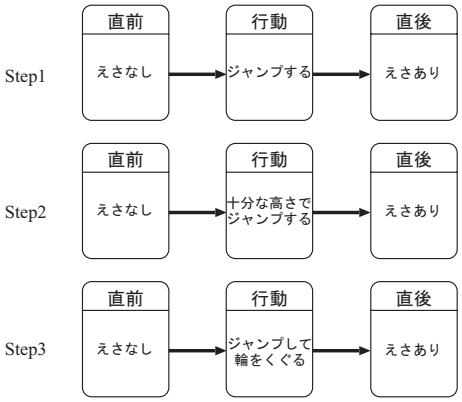


図 1. イルカにおける好子の出現による shaping

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。ここで注意する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、強化したい行動がうまく強化されない。また、与える好子の量が多くても少なくても駄目である。この例の場合、一回に与えるエサの量が多すぎる場合、イルカは行動が強化される前に満腹になってしまい、繰り返しトレーニングできない。逆に少なすぎてもイルカは報酬に対して満足できなくて再度、同じ行動をしようと思わない。

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。

3　Shaping 強化学習を用いた行動獲得手法

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、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。

これに対し、本研究室ではこれまで階層型ファジィ行動制御の上位の行動選択部に強化学習を用いる手法やマクロ状態認識ファジィルールにより学習器の入力次元数を減少させ、学習時間の短縮を図る研究を行ってきた。これまではロボット内部で知覚した情報を効率的に処理する方法のみを設計 (内部環境設計) してきたが、本研究では前述した Shaping の概念を強化学習に取り入れ、ロボッ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 (外部環境設計) を行なうことによりロボッ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移動ロボッ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移動ロボットに与える報酬関数等を人為的に変動させる。

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。Shaping を加える方法は、ロボッ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。

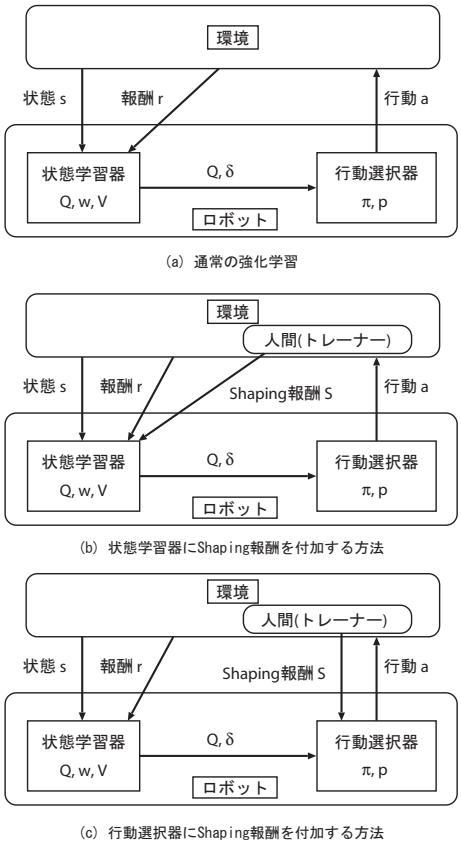


図 2. 提案する Shaping 強化学習の概念図

3.1 Q-Learning

Q-Learning に Shaping の要素を組み込む方法として 2 通りの提案をする。まず 1 つ目は Shaping の要素をサブ報酬という形で与える方法、2 つ目は Q-Table とは別に Shaping-table というものを設定して行動選択の際にその値を参照する方法である (図 2 参照)。

(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) 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違い、人間が適宜、与えるサブ報酬であると定義する。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 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) のように人間がロボッ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ることに注意されたい。

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

T : 温度定数
$S(s_t, a_t)$: 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 Profit Sharing

Profit Sharing はルール系列に対して一括で報酬を与える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。ここでは Shaping 報酬を組み込んだ式 (3) のような強化関数を提案する。

(1) 行動重みに Shaping 報酬を付加する方法 (SPS)

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 報酬を与えている方法と同様の考え方である。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、連続した行動系列を効率よく学習しようとする考え方である。

$$f(h) = \gamma^h (r_t + S(t)) \quad (3)$$

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

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)
$else\ S(t) = 0$

3.3 Actor-Critic

Actor-Critic は Actor(行動決定部) と Critic(状態評価部) が明示的に分かれているので、どちらに Shaping の要素を組み込む方がより効率的かを検証する必要がある。ここでは Actor と Critic のそれぞれに Shaping の要素を組み込んだ場合の 2 つの手法を提案する。

(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。ロボットは与えられた地点 (状態) を重要な価値があると考える。

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

δ_t : TD 誤差　　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) Actor に Shaping 報酬を付加する方法 (SAC2)

行動価値関数の更新式に Shaping 報酬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。ロボットは Shaping 報酬が与えられた行動と状態を体で覚えてその状況を記憶することに相当する。

$$p(s_t, a_t) \leftarrow p(s_t, a_t) + \beta(\delta_t + S(t)) \quad (6)$$

β : ステップサイズパラメータ
$S(t)$: Shaping 報酬
if 人間が調教を行った時　$S(t) = c$　　(c は定数)
$else\ S(t) = 0$

4　分化強化型 Shaping 強化学習を用いた行動獲得手法

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、ロボッ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。そこで、2 章で紹介した段階的に行動を強化していく分化強化を取り入れ、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。この手法の比較として、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を与える Q-Learning(以後、Sub-QL とする) を考える。DR-SQL の学習手順を以下に、アルゴリズムフローを図 3 に示す。この手順で Shaping 強化学習を行なうとロボッ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。

Step1 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=1)

Step2 ロボッ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。

Step3 サブゴール i までの学習がある程度収束してきたら学習を終了する。

Step4 サブゴール i までの学習が終了したら、次に移行する.

Step5 サブゴール i までに学習した Q-table を S-table に加算する。

Step6 Q-table を初期化して、サブゴール i を消去する。

Step7 学習終了条件 (サブゴール i がゴールと等しい) 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、**Step2** へ戻る。

5　シミュレーション実験

3 章、4 章で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。シミュレータの概観を図 4 に示す。

基本的には 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、ロボットに Shaping を行なう場合、調教者 (人間) がシミュレータ上のロボッ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。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) を考える。
- ロボットのスタート (初期位置) は空間の左下、ゴールは右上とする。
- ロボッ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。
- ロボットは上下左右、斜めに進む行動をとることができ、合計 8 つの行動を自律的に選択する。
- ロボットはグリッドの外には出ないものとし、外に出るような行動を選択した場合は再度行動選択する。
- ゴールで得られる報酬を 10 × 10 のグリッドでは 10 として、20 × 20 のグリッドでは 100 とした。

実験 1(様々な学習法を用いた Shaping 強化学習の比較)

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 4(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 の状態学習器に Shaping 報酬を付加する方法 (SQL1)
- Q-Learning の行動選択器に Shaping 報酬を付加する方法 (SQL2)
- Profit Sharing の行動重みに Shaping 報酬を付加する方法 (SPS)
- Actor-Critic の Critic に Shaping 報酬を付加する方法 (SAC1)
- Actor-Critic の Actor に Shaping 報酬を付加する方法 (SAC2)

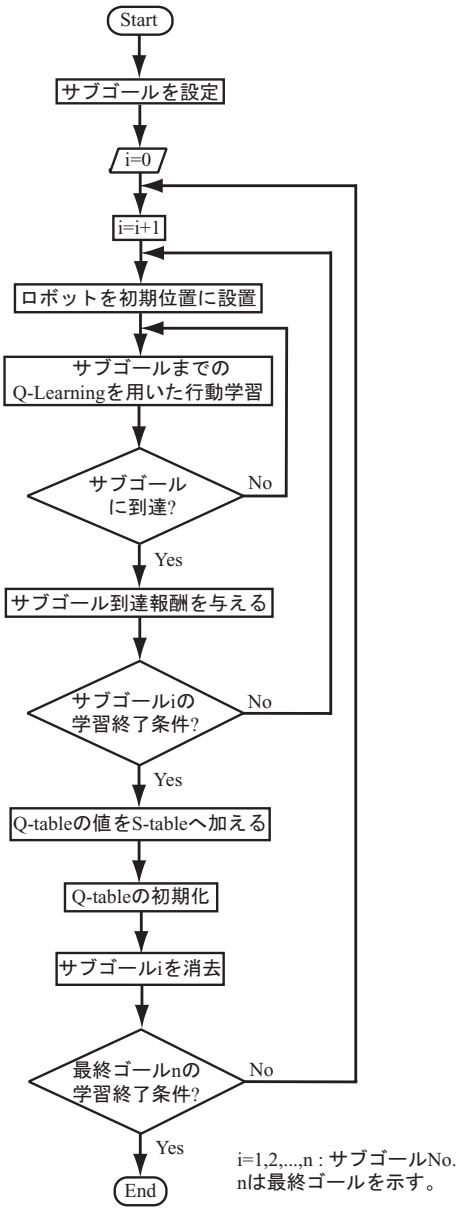
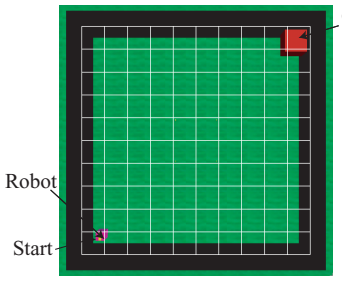
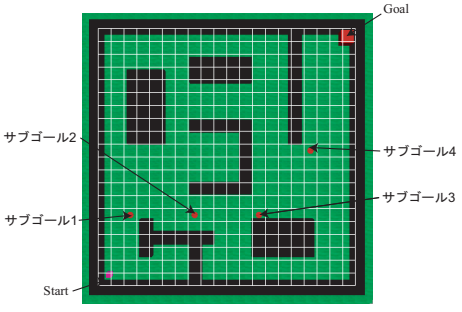


図 3. DR-SQL のアルゴリズムフロー



(a)10 × 10 グリッド:障害物なし



(b) 20 × 20 グリッド:障害物あり

図 4. シミュレータ概観

実験 2(DR-SQL の性能比較実験)

DR-SQL の性能を評価するために図 4(b) のような環境に 4 つのサブゴールを設定してシミュレーション実験を行なった。比較には通常の Q-Learning と Sub-QL(サブゴールの設定場所は DR-SQL と同様) を用いた。DR-SQL のサブゴール到達報酬は 5、Sub-QL のサブゴール到達報酬は 0.1 を与えた。

5.1 シミュレーション結果および考察

実験 1、2 の結果を図 5、6 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。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することもあった。しかしながら、この手法は環境が複雑になると Shaping 報酬を多く与える必要が出てきて、調教者の負担が大きくなるといった問題が発生する。

そこで、段階的に強化を行なっていく分化強化型 Shaping 強化学習を提案し、20 × 20 のグリッド (障害物あり) の環境下でシミュレーション実験を行なった。結果は Shaping 強化学習と比べると調教者の負担が非常に軽く、性能も良いことがわかった。今回は簡易的に各サブゴールで得られるサブゴール到達報酬の量を一定で行なったり、サブゴールの終了判定を十分吟味していないなど、実際の動物などの調教で重要な点を深く議論していないので、今後、検討していく必要がある。

6 結 言

本研究では Shaping 強化学習を用いた自律移動ロボッ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 をロボッ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強化学習を用い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。

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)

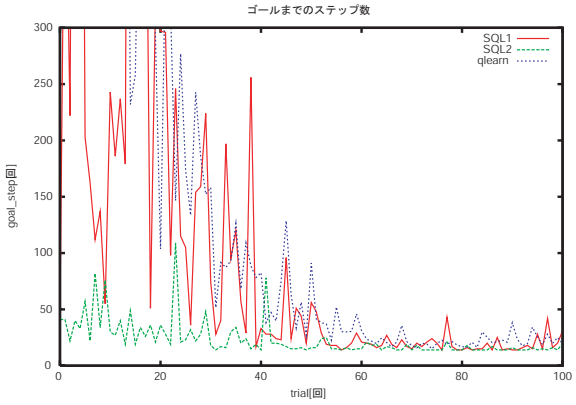
[2] H. Ishiguro, R. Sato, T. Ishida, ”Robot oriented state spaceconstruction, ” *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'96)*, pp.1496-1501 (1996)

[3] 野田彰一, 浅田稔, 倹積田健, 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)

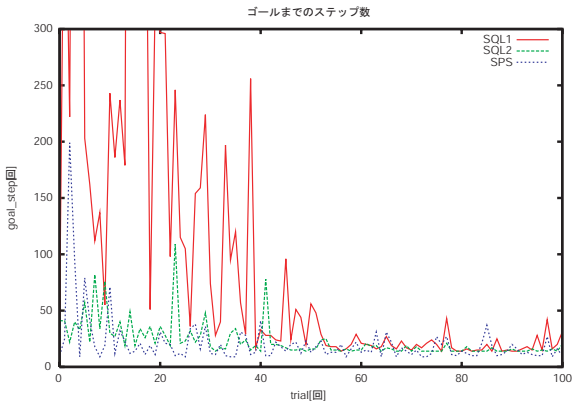
[4] 杉山尚子, 島宗理, 佐藤方哉, リチャード・W・マロット, マリア・E・マロット, ”行動分析学入門, ” 産業図書 (1998)

[5] 花香敏, 前田陽一郎, ”自律移動ロボットの戦略獲得のためのファジィ状態分割型強化学習, ” 第 23 回日本ロボット学会学術講演会, CD-ROM, 3E12 (2005)

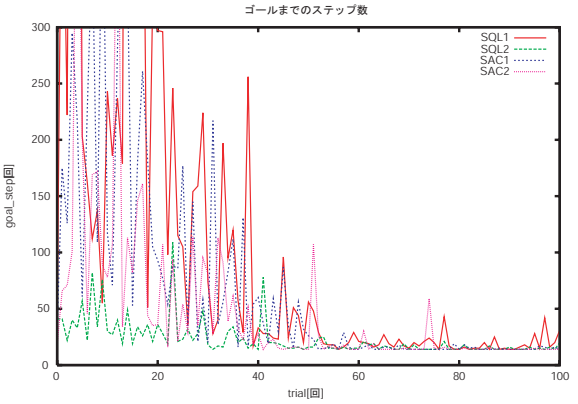
[6] 花香敏, 前田陽一郎, ”Shaping 強化学習を用いた自律移動ロボットの行動獲得, ” 第 18 回ファジィ・コンピューティング研究部会ワークショップ, 05-18 (2006)



(a) SQL



(b) SPS



(c) SAC

図 5. 実験 1 の結果

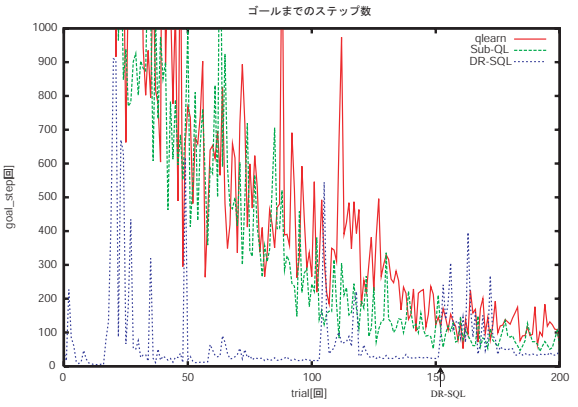


図 6. 実験 2 の結果