1D3-05

強化学習における自己組織化マップを用いた状態空間の自律的構成法

A system of autonomous state space construction with the self-organizing map in reinforcement learning

岩崎 秀樹[*1] 末田 直道[*1]
Hideki Iwasaki Naomitchi Sueda

[*1]大分大学工学部
School of Engineering, Oita University

A state space construction is a very important problem in the application of reinforcement learning to real tasks. A system for state space construction with self-organizing map is proposed in this paper. In this system, the agent constructs state space from its own experience autonomously. In the experimentations, the system verifies the agent's ability to construct a suitable state space from any unknown situation. Subsequently, it can improve ability and steadiness of learning, and robustness to noise. Furthermore, it is capable of reconstructing the state space to best fit any change in the environment.

1. はじめに

近年，自律的にタスクを遂行するエージェントやロボットの実現が強く望まれている．その中で，環境との相互作用と通じて適応的な学習が可能な強化学習法が注目されている．

しかし，実タスクにおいて，状態量は位置座標や速度といった連続値ベクトルで与えられることが多いが，強化学習の枠組みにおいて状態は離散的に表現される．そのため，知覚情報をエージェント内部でどのように表現するかという状態空間の構成問題が生じる．その際，より少ない状態数で環境の特徴を捉えた状態空間を構成することが望ましい．状態空間の構成は学習の性能に大きく影響する重要な問題である．

本研究では，自己組織化マップを用いた状態空間の構成法を提案し，エージェントによる自律的な状態空間の構成を目的とする．そこで，連続平面上を障害物を回避しつつ目標地点で停止するというタスクを想定し，シミュレータによる実験を行う．そして静的環境やノイズを含む環境および変化的な環境による実験を行い，提案手法の未知環境への適応性，ノイズに対する頑健性，また環境の変化に対する学習の追従性を検証する．

2. 強化学習

強化学習 [Sutton98] とは，環境との相互作用と通じて最適な行動戦略を獲得する機械学習のひとつである．

強化学習の代表的手法として，Q-Learning[Watkins92] がある．Q-Learning では，状態 s において政策 π にしたがって行動 a を選択し，環境が行動 a を実行する．そして，環境から与えられる報酬 r，次状態 s' に対して，次式を用いて状態行動価値 $Q(s,a)$ を更新する．

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$$

ここで，α は学習率，γ は探査率である．

提案手法における学習方法として，Q-Learning を用いる．

連絡先: 岩崎秀樹，大分大学大学院工学研究科，〒 860-157 大分県大分市大字旦野原 700 番地，TEL:097-554-7866, FAX:097-554-7886，E-MAIL:hide0128@csis.oita-u.ac.jp

3. 自己組織化マップ

自己組織化マップ (以下，SOM)[Kohonen90][Kohonen96] は，教師なし学習により，入力データの位相関係を保持した特徴マップを形成する．本研究では，エージェントが得られる知覚情報から，エージェント内部の状態空間を構築するために SOM を用いる．

SOM は，入力層と出力層の 2 層構造からなるニューラルネットワークであり，入力層のノード (入力ノード) と出力層のノード (出力ノード) は完全結合している．また，各出力ノードは入力ベクトルと同次元の荷重ベクトル $\mathbf{w}$ を持つ．入力層に入力ベクトル $\mathbf{x}$ を入力すると，出力層は入力ベクトルとの距離が最小である荷重ベクトルを持つ出力ノード (勝者ノード)c を決定する．そして，勝者ノード c とその近傍ノード i に対して，荷重ベクトル $\mathbf{w}_i$ を次式によって更新する．

$$\mathbf{w}_i^{\mathrm{new}} = (1-\alpha(t))\, h_{ic}(t)\, \mathbf{w}_i^{\mathrm{old}} + \alpha(t)\, \mathbf{x}$$

$$h_{ci}(t) = \exp\left(-\frac{||\mathbf{r}_c - \mathbf{r}_i||}{2\sigma^2(t)}\right)$$

ここで，t は時刻，$\alpha(t)$ は学習係数 $(0 < \alpha(t) \leq 1)$ である．また，$\sigma(t)$ は近傍の広がりを示し，$\alpha(t)$ と $\sigma(t)$ は学習の進展に伴い小さくする．

学習の結果得られる荷重ベクトルの分布は，入力空間上の入力ベクトルの分布を近似する．つまり，入力ベクトルが一様に分布する場合は，SOM の荷重ベクトルも一様な広がりを持つようになり，入力ベクトルの分布に偏りがある場合は，荷重ベクトルもその偏りを持つ分布となる．

4. 移動停止問題

状態空間の構成問題を含む問題例として，本研究では移動停止問題を想定し，シミュレータ (図 1) を用いた実験を行う．移動停止問題の構成は以下に示す通りである．

＜環境＞

・移動領域：縦 200，横 200 の連続平面

・ゴールエリア：中心座標 (x,y)，半径 r で定義される円．エージェントはこのエリア内で停止することを目標とする

・障害物：障害物は左上の頂点座標 (x,y) と横幅，縦幅で定義される長方形の物体である．エージェントは障害物に衝突すると罰が与えられる．

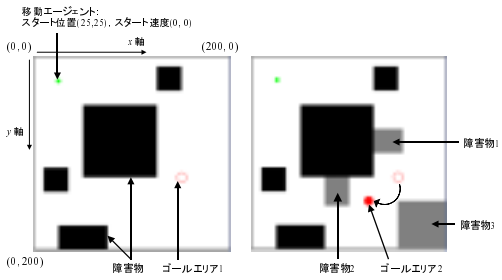

図 1: 実験用シミュレータ (左：実験 1(静的環境)，右：実験 3(変化的環境))

・壁：移動領域の周囲は壁であり，障害物と同様にエージェントは壁に衝突すると罰が与えられる．
・エージェントのスタート位置・速度：各エピソードごとにエージェントの位置，速度をこの値に初期化する
・エージェント情報：環境は現時点でのエージェントの正確な位置 $\mathbf{pos}$: (pos_x, pos_y)，速度 $\mathbf{vel}$: (vel_x, vel_y) を有する．各ステップ毎，エージェントから加速度 $\mathbf{accel}$: $(accel_x, accel_y)$ を受け取り，以下の式により速度，位置の更新を行う．

$\mathbf{vel} \leftarrow \mathbf{vel} + \mathbf{accel}$，$\mathbf{pos} \leftarrow \mathbf{pos} + \mathbf{vel}$

壁や障害物との衝突がない場合は終了．
以下，位置 $(point_x, point_y)$ で衝突した場合，
横 (左右) から衝突した場合：$vel_x \leftarrow 0$，$pos_x \leftarrow point_x$
縦 (上下) から衝突した場合：$vel_x \leftarrow 0$，$pos_y \leftarrow point_y$
・エージェントへの知覚情報の提供：環境は知覚情報として，エージェントの位置，速度情報を提供する．

＜エージェント＞
・知覚情報：速度情報 (v_x, v_y)，位置情報 (p_x, p_y)：これは各ステップ毎に環境から与えられる情報である．
・行動：加速度 (a_x, a_y) $(-1 \leq a_x, a_y \leq 1)$：これは各ステップ毎に環境に送る行動である．実験において選択可能な行動として，図 2(左) に示す 25 個の加速度ベクトルを持つ．

＜タスク＞
・目標：エージェントはゴールエリア内で停止することを目標とする．ただし停止条件は，$1.0 < |vel_x|, |vel_y|$
・報酬・罰：エージェントは各ステップにおいて，環境から次の報酬 (罰) を受け取る．目標達成に対して：100，衝突に対して：-10，移動コスト：-1(毎ステップ)

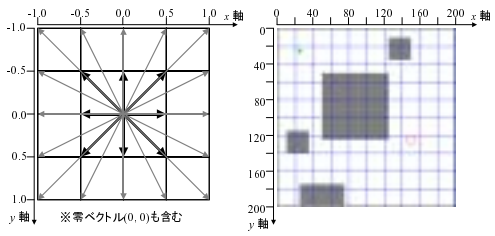

図 2: 左：選択可能な加速度ベクトル, 右：グリッドによる位置情報の構成

5. 提案手法

提案手法の概要を図 3 に示す．エージェントは行動戦略の学習のための学習機構と，環境から与えられる知覚情報から状態空間を構成するための状態構成機構を有する．

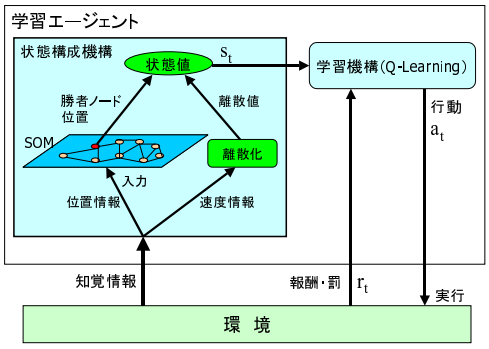

図 3: 提案手法の概要

5.1 状態構成機構

エージェントが環境から与えられる知覚情報は位置と速度情報であり．それぞれ 2 次元の連続ベクトルである．そのため，知覚情報を学習機構に用いるためには，エージェントは知覚情報に対して離散的な状態空間を構成する必要がある．このような状態空間を構成するのが状態構成機構である．状態構成機構は，位置情報と速度情報に関してそれぞれ以下の構成を用いる．そして，それざれから得られる離散値から離散的な状態値を生成し，学習機構に提供する．

＜位置情報の構成＞
状態構成機構は位置情報に関して，SOM を用いて状態空間を自律的に構成する．

移動停止問題に対して，位置情報に関する状態空間の構成を考えるとき，障害物内部の空間は，エージェントが経験することのない空間であり，この空間を状態空間に割り当てることは無駄である．一方で，ゴールエリア周辺では，エージェントはゴールエリア内で止まるために，より細やかな制御が必要となるため，位置情報を細かく構成することが望ましい．また，ゴールから遠く，障害物が少ない領域では衝突の危険性も低く，ある程度粗い制御でもタスクの達成に影響せず，位置情報の分割も粗くしても学習に大きく影響しないと考えられる．つまり，状態空間の構成では，タスクにおいて重要視すべき領域は細かく，学習に大きく影響しない領域は粗く構成したほうが，より効率的な学習を行うことができる．しかし，設計者が事前に適切な状態空間を構成することは困難であり，エージェントが自律的に適切な状態空間を構成でき，かつ環境の変化に対して柔軟に状態空間を適応させることが望ましい．

そこで，提案手法では自己組織化マップを用いて状態空間を構成し，エージェントが状態空間に関して未知の状況から自律的に適切な状態空間を構成する．また，SOM に関する各パラメータの設定値を図 4(左下) に示す．ここで，学習係数と近傍関数の最小値の制限は SOM のの適応性を維持するものであり，環境の変化に適応するための重要なパラメータである．

＜速度情報の構成＞
速度情報に関しては x, y 軸ともに図 4(右) で示すような構成を用い，離散化する．これは，エージェントに事前に与えられるものであり，固定的である．

5.2 学習機構

学習機構では，環境から与えられる報酬 (罰) と，状態構成機構を介して与えられる状態値をもとに行動選択を行う．ま

た，一連の試行を通して，各状態における最適な行動戦略を学習する．学習アルゴリズムとしては Q-Learning を用いる．方策 π として ϵ-greedy 方策を用いる．これは，ϵ の確率で一様に行動を選択し，それ以外では最大の $Q(s,a)$ を持つ行動を選択するものである．提案手法での学習に関するパラメータは図 4(左上) で示す通りである．

6. 実験

シミュレータを用いて，以下の 3 つの実験を行う．そして，位置情報に対してグリッド状 (図 2 : 右) に状態空間を構成する手法との比較を行い，提案手法の性能を検証する．グリッドによる構成では，位置情報を x, y 軸ともに等間隔に分割し，各ブロックの範囲内をひとつの (位置情報に関する) 状態とする．

6.1 実験内容

＜実験 1：静的な環境＞

実験 1 では，まず位置情報に関して全くの未知の状態から，適切な状態空間を構成することができるかという環境への自律的適応性を検証するために，シミュレータを図 1(左) のように設定した環境における実験を行う．また，エージェントの知覚情報に関して，エージェントは環境から現時点の正確な位置情報，速度情報が与えられるものとする．

＜実験 2：ノイズを含む環境＞

実問題において，エージェントが知覚できる情報にはノイズが含まれていたり，環境に予測不能な未知のパラメータが含まれたりすることが多い．実験 2 では知覚情報のうち，位置情報に対してノイズを含む実験を行う．実験 2 の環境の設定について，エージェントの知覚情報以外は実験 1 と同じものとする．エージェントの知覚情報は，位置情報に関して，ノイズとして平均 0，分散 1 の正規乱数を付加された位置が与えられる．速度情報は，正確な速度が与えられるものとする．

＜実験 3：変化的な環境＞

実問題では，環境が時間に伴って変化する場合も考えられる．例えば，移動ロボットでは移動空間の障害物の配置が変わったり，目標地点が変わったりする状況が考えられる．この際，エージェントが自律的に環境の変化に対して，追従するように学習を進めることができるのが望ましい．そこで，実験 3 ではある学習の途中で，障害物の追加とゴールエリアの移動を行い，提案手法の環境の変化に対する学習の追従性を検証する．実験 3 での環境 (図 1 右) の設定は，ゴールエリアと障害物以外に関して実験 1 と同様である．ゴールエリアは初期位置は実験 1 と同様であるが，50,000 エピソード終了後に，ゴールエリア 2 に変更する．障害物 1,2,3 はそれぞれ，30,000 エピソード，50,000 エピソード，700,000 エピソード終了後に追加される．

パラメータ	値
学習率 α	0.1
割引率 γ	0.05
探査率 ε	0.95

学習機構の設定値

パラメータ	値	最小値
出力ノード行数 Row	10	
出力ノード列数 Col	10	
学習係数 α(t)	$1\times(0.99995)^t$	0.0001
近傍の広がり σ(t)	$10\times(0.99999)^t$	0.5

SOMの設定値

速度	離散値
vel< -8	0
-8≦vel<-6	1
-6≦vel<-4	2
-4≦vel<-2	3
-2≦vel< 0	4
0≦vel< 2	5
2≦vel< 4	6
4≦vel< 6	7
6≦vel< 8	8
8≦ vel	9

速度情報の離散化

図 4: 各設定値

6.2 実験結果と考察

＜実験 1：静的な環境＞

実験 1 における，位置情報に関する SOM マップの構成の様子を図 5 に示す．エージェント内部の SOM マップ (出力層の荷重ベクトルによって形成されるマップ) は，初期化によってランダムに形成されている．エージェントはまず，学習序盤においてこのもつれたマップを整えるように，位相保持マップを形成する (1 エピソード (序盤，終盤) の図)．100 エピソードまで学習が進むと，ほぼ均等なグリッド状に SOM マップが形成される．ここで注目すべきなのが，障害物が存在する領域では，SOM マップの格子が粗く配置されていることである．これは，障害物内部はエージェントが決して経験することのない状況であり，SOM に入力されないため，SOM の特性により SOM マップも粗くなるためである．一方，学習の進展に伴い，エージェントはゴールに対し似たような経路を選択するようになる．そのため，その経路とゴールエリア周辺が SOM に入力されることが，他の領域に比べ多くなる．そのため，学習後半 (10,000 と 100,000 エピソードの図) では，徐々にゴールまでの経路とゴールエリア周辺の SOM マップが細かく形成されていく．このような段階を経て，エージェントは位置情報に関して，全くの未知の状況から目標達成のために適した状態空間の構成をすることができる．

次に，目標達成に要したステップ数の 100 エピソード毎の平均を表すグラフを図 7(上段) に示す．これを見ると，提案手法はグリッドによる位置情報の構成よりも，学習の収束も早く，学習によって得られる行動戦略の性能も良いことがわかる．また，グリッドによる構成では，その学習に揺らぎが多く見られるが，提案手法では学習の揺らぎは非常に少ない．

つまり，SOM を状態空間の構成に利用することで，適切な状態空間を構成をすることができるため，学習の性能と安定性の向上を図ることができたといえる．

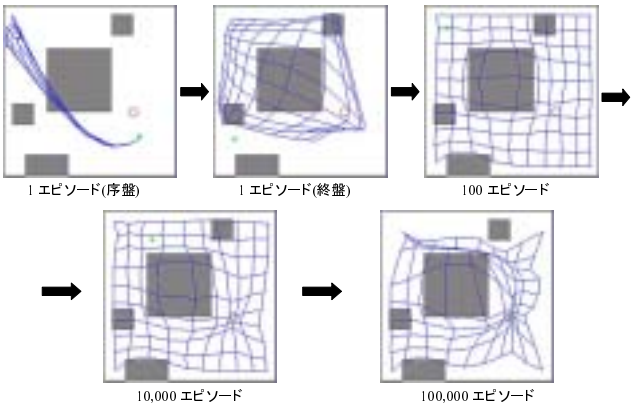

図 5: 位置情報に関する SOM マップの構成

＜実験 2：ノイズを含む環境＞

実験 2 に対する平均ステップ数を図 7(中段) に示す．ここで，グリッドによる構成では実験 1 に比べ学習の揺らぎが多く，またその揺らぎの振幅も大きくなっていることがわかる．ノイズを含む環境においてグリッドによる構成では，そのノイズの影響を強く受け，学習の性能は著しく低下している．これに対し，提案手法では実験 1 の結果に比べ，学習の収束値は若干高いものの，収束後は安定した性能を示している．つまり，提案手法

によりノイズに対する学習の頑健性を向上することができる．

＜実験 3：変化的な環境＞
実験 3 における，位置情報に関する SOM マップの構成の様子を図 6 に示す．エージェントは，ゴールエリアが移動するまでは実験 1 の結果 (図 5) と同じように障害物の領域は粗く，ゴールまでの経路とゴールエリア周辺は細かくなるように SOM マップ (50,000 エピソードの図) を形成する．50,000 エピソード終了後，ゴールエリアは移動するが，この時は当然，移動後のゴールエリア周辺の SOM マップは粗い．しかし，学習の進展に伴って，SOM マップは移動後のゴールエリア周辺を細かく構成するように変形していく．このように提案手法では環境の変化に追従するように状態空間の再構成することができる．

次に，実験 3 に対する平均ステップ数を図 7(下段) に示す．ゴールエリアの移動 (50,000 エピソード) までは実験 1 と同様に，提案手法はグリッドによる構成よりも安定してよい性能を示している．そして，ゴールエリア移動後に対しても，上で述べた状態空間の再構成により，移動前とほぼ同様の安定した性能を発揮することができている．

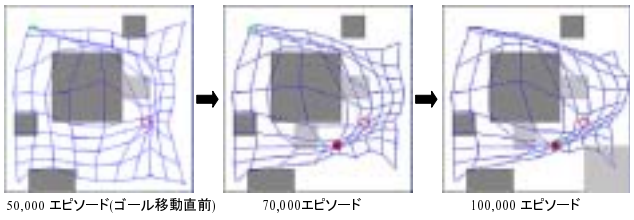

図 6: SOM による位置情報の構成 (環境の変化への追従)

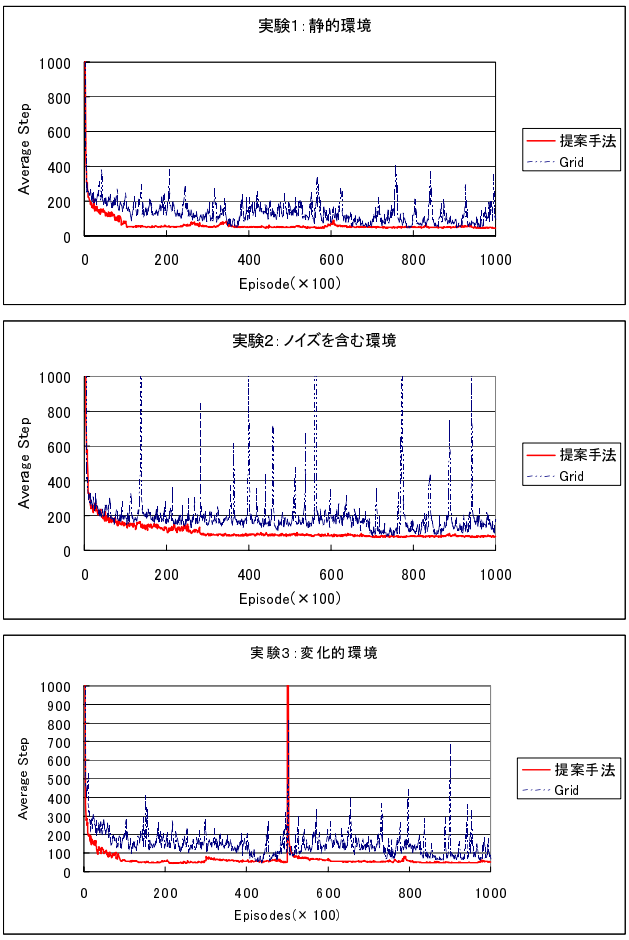

図 7: 実験結果

7. まとめと今後の課題

7.1 まとめ

本研究では，強化学習を実タスクに適用する際に，重要な要素となる状態空間の構成問題に着目した．特に未知環境や動的な環境において，エージェントが経験を通して自律的に適切な状態空間を構成することを目指し，自己組織化マップを用いた構成法を提案した．

実験では，状態空間の構成問題を含むタスクとして，移動停止問題を想定し，提案手法の評価を行った．この実験で提案手法により，エージェントは位置情報に関して，未知の状況から自律的に適切な状態空間を構成することができ，学習によって得られる行動戦略の性能を向上させることができることを示した．また，適切な状態空間の構成ができることにより，ノイズのような未知のパラメータを含む環境に対する頑健性を向上することができる．そして，学習の途中で環境が変化するような場合においても，エージェントは自律的にその変化に追従するように状態空間を再構成するため，環境の変化後も学習の性能を維持することが可能となる．

7.2 今後の課題

提案手法では，環境から与えられる知覚情報のうち位置情報に関して，エージェントの自律的構成を行い，速度情報に関しては，固定的な構成を用いた．また行動空間に関しても，事前にその構成をエージェントに与えて，学習を行った．しかし，速度情報，行動空間に関しても，エージェントが自律的に構成する望ましい．そこで，提案手法を拡張し，状態空間および行動空間を総括的に自律的構成をできるようにすることが一つの課題である．

また，提案手法で用いた学習機構や SOM の設定値は経験的に得たものであり，必ずしも最適な値であるとは限らない．そのため，これらの設定値の検討をする必要がある．また，提案手法では，学習方法として Q-Learning を用いたが，他の学習方法 (Sarsa や Profit Sharing) を組み込んだ場合での，学習性能を検証する必要もある．

参考文献

[Sutton98] Richard S. Sutton, Andrew G. Barto; Reinforcement Learning: 三上 貞芳, 皆川 雅章 共訳: 強化学習: 森北出版 (2000)

[Watkins92] Watkins, C. J. C. H. and Dayan, P. :Technical Note: Q-Learning, Machine Learning 8, pp. 279–292 (1992).

[Kohonen90] T.Kohonen: The Self-organizing Map: Proc. of the IEEE, pp. 1464-1480(1990)

[Kohonen96] T.Kohonen; SELF-ORGANIZING MAPS; 徳高平蔵 岸田悟 藤村喜久朗 訳: 自己組織化マップ: シュプリンガー・フェアラーク東京 (1996)