6L-5

強化学習における階層型 SOM を用いた状態空間の適応的構成法

岩崎 秀樹† 末田 直道‡

大分大学大学院工学研究科知能情報システム工学専攻† 大分大学工学部‡

1. はじめに

強化学習を実問題に適用する場合，状態量は位置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要素である．本研究では，自己組織化マップ（SOM）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，移動停止問題に対し実験を行い．提案手法を用いることで適切な状態空間の構成を実現し，学習性能（収束性能，獲得戦略の性能，安定性）が向上することを示す．

2. 利用技術

2.1 強化学習

強化学習[Sutton98a]とは，環境との相互作用を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．提案手法では，その代表的なアルゴリズムである Q-Learning[Watkins92]を用いる．Q-Learning において，一般に状態は離散的に表現される．

2.2 自己組織化マップ（SOM）

SOM（Self Organizing Map）[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．SOM は，入力層と出力層からなるニューラルネットワークである．各出力ノードは入力ベクトルと同次元の荷重ベクトルを持つ．入力ベクトルが与えられると，そのベクトルと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）を決定する．そして，荷重ベクトルを更新することで学習が進展する．学習により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．

A system of autonomous state-space construction with a hierarchical SOM in reinforcement Learning
Hideki Iwasaki †, Naomichi Sueda ‡
† Graduate School of engineering, Oita University
‡ Department of Computer Science and Intelligent Systems,

3. 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左)を用いた実験を行う．移動停止問題において，エージェントは障害物の配置された連続平面上を移動し，ゴールエリア内で停止することを目標とする．エージェントは，各ステップ毎に環境から知覚情報として，位置・速度情報（ともに２次元連続値ベクトル）を与えられ，行動として，図１右で示される加速度ベクトルの一つを選択し，環境に返す．環境はそれを受けて，エージェントの位置と速度を更新する．また，エージェントは，１ステップ毎に移動コスト，壁や障害物との衝突に対し罰，目標達成に対し報酬が与えられる．

移動停止問題において，重要なことは位置によって適切な状態空間の構成が異なることである．ゴールエリア付近では，ゴールエリア内で停止するために，細やかな制御が必要であり，位置・速度情報ともに状態空間も細かく特徴付ける必要がある．一方，ゴールから遠い領域では，より早くゴールに向かうために速い速度領域の状態空間を重要視する必要がある．

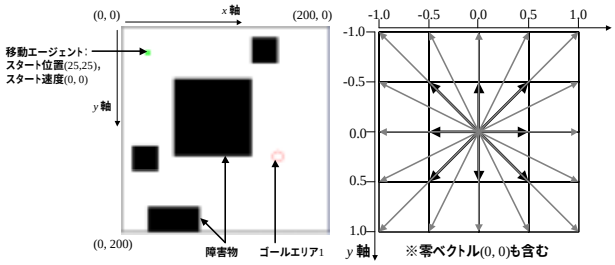

図１ 左：実験環境，右：行動ベクトル

4. 階層型 SOM を用いた状態空間の構成法

提案手法の概要を図２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．

学習機構は，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とともに，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．提案手法において学習アルゴリズムとしては Q-Learning を用い，方策としてεグリーディ方策を用いる．

状態構成機構は，環境から連続値ベクトルで

与えられる知覚情報を，離散的な状態空間として構成するものである．状態構成機構には位置情報に対する状態空間を構成する SOM（**posSOM**）があり，その posSOM の各ノードに対して，それぞれ速度情報に対する状態空間を構成する SOM(**velSOM**)を１つずつ持つという階層構造になっている．そして，velSOM のそれぞれのノードが１つの状態を表現する．

状態構成機構では，環境から知覚情報が与えられると，まず posSOM に位置情報が入力され勝者ノードを決定する．そして，その勝者ノードに対応する velSOM に速度情報を入力し，それによって決定される勝者ノードを状態として学習機構に送る．

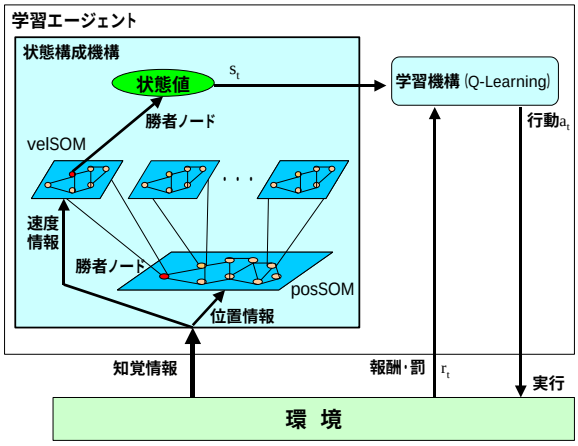

図２ 提案手法

5. 実験・結果

5.1 実験

移動停止問題に対し，提案手法と位置・速度情報をそれぞれ等分割して状態空間を構成する手法（グリッド１，２）との比較を行い，提案手法の学習性能（収束性能，獲得戦略の性能，安定性）を検証する．提案手法とグリッド分割の状態数は，以下の通りである．

	位置情報	速度情報	状態数
提案手法	10×10	5×5	２５００
グリッド１	10×10	10×10	１００００
グリッド２	15×15	10×10	２２５００

5.2 結果・考察

この実験の結果を図３に示す．ここで，横軸は試行数，縦軸は目標達成に要したステップ数の１００試行ごとの平均であり，ステップ数が低いほど獲得した戦略性能が高いことを示す．これを見ると，学習序盤の収束速度はグリッド１，２に比べ若干遅いものの，最終的にほぼ収束するのに要する試行数は，早いことがわかる．また，グリッド１では，グラフに揺らぎが多く見られ，学習が安定してないのに対し．提案手法は，学習後半ではグリッド２同様に，安定して高い性能を示していることがわかる．

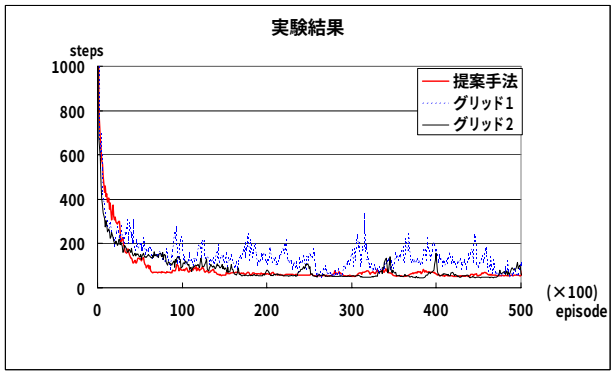

図３ 実験結果

6. 終わりに

本研究では，強化学習における状態空間の構成問題に着目し，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，階層型 SOM を用いた構成法を提案した．実験では，移動停止問題を想定し，提案手法により，エージェントはより少ない状態数で自律的に適切な状態空間を構成することができ，学習性能を向上することを示した．

また，提案手法では学習アルゴリズムとして Q-Learning を用い，行動は事前に離散化された加速度ベクトルから選択するものであった．今後の課題として，actor-critic[Sutton98b]や Fuzzy Q-Learning[Horiuchi99]といった連続行動出力に適用可能なアルゴリズムを組み合わせ，連続行動空間に対する適応性について検証する必要がある．

参考文献

[Sutton98a] 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)

[Watkings92] Watkins, C. J. C. H. and Dayan, P. :Technical Note: Q-Learning, Machine Learning 8, pp. 279--292 (1992).

[Kohonen96] T.Kohonen: The Self organizing Map:Proc. of the IEEE,pp.1464-1480(1990)

[Sutton98b] Sutton, R. S. & Barto, A.: Reinforcement Learning: An Introduction, A Bradford Book, The MIT Press (1998).

[Horiuchi99] 堀内匡,藤野昭典,片井修,椹木哲夫：連続値入出力を扱うファジィ内挿型 Q-learning の提案,計測自動制御学会論文集, Vol.35, No.2, pp.271--279 (1999).