強化学習における状態空間の自動決定

池田善治 *

1　はじめに

近年，ロボットの発達とともに機械学習の必要性が向上してきた．ロボットに動作を設計する際に全ての情報を予め教えるのは不可能に近いためである．

本研究では，機械学習の一つである強化学習に着目した．強化学習というのは，数値化された報酬を最大にすべきかを学習するものである．学習者（エージェント）はどの行動を取るべきか教えられず試行錯誤を繰り返しながら，たくさんの報酬に結びつく行動を見つけ出す．このような特徴のため，未知の問題でも，試行錯誤を繰り返して自身の経験から学ぶことが出来る学習である．[1][2]

強化学習で使用する項目は，学習時間やメモリの問題等の面から，抽象化を行い必要最低限の数に絞って設定する．しかし実機のロボットはセンサで読み取った値（実データ）を扱っている．実データはセンサから読み取る値であるため，スカラー値で表されており，強化学習で用いるために抽象化はされていない．このため実機に強化学習を適用する際，実データを直接使うのは難しいといった問題がある．実データを扱う問題の中で，最も困難な問題の一つに状態の構成がある．状態とはロボットが認識することが出来る環境の状態のことをである．エージェントが強化学習を行う時には，状態一つ一つについてどのような行動を取ればより多くの報酬が得られるのか学習を行う．このため全ての実データを状態として学習を行うと状態の数が大量に出来てしまい，学習効率の低下を招く．よって実データを扱えるようにするために従来の手法では，設計者が予め実データから状態を特定できるように設計を行い，学習を行う時には，実データから状態を決定して強化学習を用いる．

しかし設計の段階で状態の設定を行うためには，学習を行う環境をある程度知る必要がある．このため環境の推測が難しい問題では状態の設定が行えない．また設計者が設定した状態は必ずしも環境に適した状態の設定になっているとは限らない．これらが原因で学習が正しく行えないことが考えられる．

本研究では，この問題を解決するためにエージェントが自動で状態の設定（状態学習）をして，タスクに対する行動学習を行う手法を提案する．この手法を用いることで学習を行う環境ごとに状態を設定するため，学習を行うたびに適した状態の設定を行うことが出来る．

2　提案システム

提案システムでは状態の設定基準を学習することで問題を解決する．今回の手法では，状態はセンサのレンジを利用して作成する．具体的にはレンジの範囲内で状態の設定基準となる値の範囲を算出して状態学習を行う（図 1）．

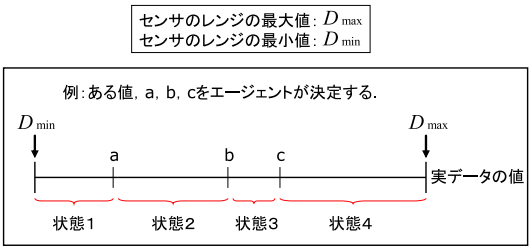


図 1: 状態設定の例

状態学習の方法は値の範囲を分割また融合することで実現する．分割とはある状態を決められた基準で複数の状態に分けることである．今回の検証では状態を中間地点で 2 分割する．融合とは隣り合わせにある 2 つの状態を繋げることである．分割を行うことにより，細かな範囲で状態の設定を行うことが出来る．また同じ状態と見なしてよい範囲を融合することで，状態の数を必要最低限に設定することが出来る．

つまり分割と融合を繰り返すことで，値の範囲の設定を繰り返し，より適した状態の設定を見つけるように状態学習を行うことが可能となる．提案手法の流れは（図 2）のようになる．

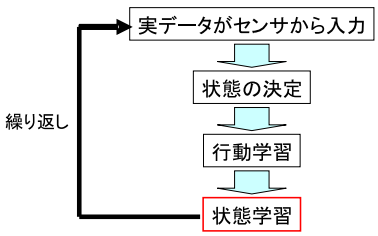


図 2: 提案手法の流れ

*室蘭工業大学 情報工学科 認知ロボティクス研究室

3　実験

本研究の実験では，まず提案システムで適切に状態の設定が出来るかどうかを検証した．次に従来の手法と比較を行い，学習効率にどのような差が出るのかを検証した．さらに実際の問題に適用することを考え，ダムの放水問題に提案手法を適用して検証を行った．ここでは，提案システムの状態の設定結果と従来手法と比較した結果を紹介する．検証はシミュレーションで行い，予めエージェントに与えるデータを作成している．ここで紹介する実験ではデータに対して適切な行動を取ると，報酬を得ることができるという問題に対して，提案システムではどのように状態が設定されるかを確認する．状態の作成を確認した後，提案システムと従来の手法の学習効率を比較する．実データ $D(k)$ と行動による報酬の関係を表 1 に示す．

表 1: 報酬の設定

	行動 A	行動 B	行動 C
$0.65 \leq D(k) \leq 1$	1	0	0
$0.35 \leq D(k) \leq 0.65$	0	1	0
$0 \leq D(k) \leq 0.35$	0	0	1

実験を行った結果を（図 3(a)）（図 3(b)）（図 3(c)）に示す．これらの図では，それぞれの行動ごとの作成された状態の分布を表している．横軸は作成したデータの範囲（実データ）であり，縦軸はその状態の時の報酬の期待値を表す．これらの結果を見ると，それぞれの行動ごとに報酬が得られる実データの時に報酬の期待値が高い状態が作成されていることがわかる．状態の設定は報酬が得られる状態とそうではない状態で明確に区別されている．

次に従来手法と提案システムを報酬の入手率で比較を行う．従来手法では，報酬の設定と同じように範囲を指定して学習を行うシステム 1 と，報酬の設定とは異なる範囲を設定して学習を行うシステム 2 を用いた．また学習を行っていないランダムに行動を選択するも比較対象にする．

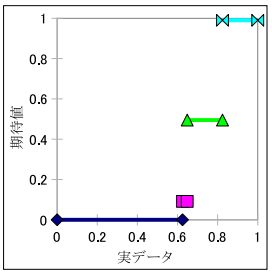
従来のシステムと提案システムの比較の結果を（表 2）に示す．結果を見ると，システム 1 では高い報酬入手率を示している．またシステム 2 では，ランダム選択よりは報酬入手率が高く学習を行っていることがわかるが，システム 1 よりも報酬入手率が低い．このため学習の効率が落ちていることが分かる．一方，提案システムではシステム 1 と同等の報酬入手率を示している．これは提案システムがシステム 1 のように適切に状態を設定して学習を行っていることを示している．

表 2: 報酬の入手量

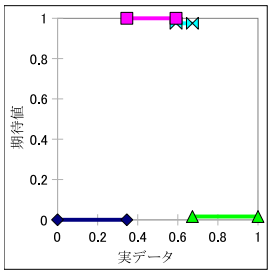
	報酬入手率（%）
提案システム	91.92
システム 1	93.34
システム 2	74.72
ランダム行動	33.65

4　まとめ

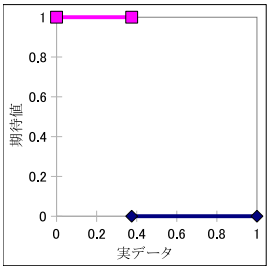
実験結果より，提案システムを使用することで状態の設定を自動で行えることを示した．しかし提案手法では時系列が関係するタスクに対応できない等問題があるので対応できるように改良する必要がある．



(a) 行動 A の状態の分布



(b) 行動 B の状態の分布



(c) 行動 C の状態の分布

図 3: 行動ごとの状態の分布

参考文献

[1] Richard S. Sutton and Andrew G.Barto，強化学習，森北出版株式会社

[2] 木村 元，宮崎 和光，小林 重信，強化学習システムの設計指針，計測と制御，Vol.38 No.10，October 1999.