2R-9

連続状態空間における状態クラスタを用いた強化学習

小野寺　道寛†　　鈴木　輝彦‡　　太原　育夫‡
†東京理科大学大学院理工学研究科情報科学専攻　　‡東京理科大学理工学部情報科学科

1　はじめに

実世界を動き回る自律ロボットには，外界の変化に対応して行動することが要求される．近年，このような適応的且つ反射的な行動をロボット自身や環境のモデルを前提とせずに獲得する手法として強化学習が注目されている [1]．強化学習を用いることによって，エージェントは得られる報酬の値を最大にする行動系列を獲得することが可能である．しかし，連続状態空間上では複数の行動系列の獲得が求められることが少なくない．従来の強化学習アルゴリズムでは単一の行動系列しか獲得できないので，そのようなタスクをエージェントに遂行させることは困難である．

本研究では，この問題点を解決するために**状態クラスタ**というものを利用した強化学習アルゴリズムを提案する．そして，障害物が存在する目的地到達タスクを用いた実験において本提案アルゴリズムの有効性を検証する．

2　状態クラスタを用いた強化学習

状態クラスタとは，必要とされる行動系列が同一で且つ状態間の距離が近い状態をひとまとめにしたものである．本学習アルゴリズムでは，そのクラスタ毎に 1 つの学習器を与えることによって，1 つのクラスタにつき 1 つの行動系列を獲得できるようにしている．これにより，複数種類の行動系列の獲得が可能となり，エージェントに連続状態空間上のタスクを遂行させることが可能になると考えた．また，タスク環境はエージェントにとって未知のものと仮定しているため，状態クラスタは動的に自動で構築させることにする．

なお，本アルゴリズムではタスク環境の状態によって必要とされる行動系列は変わるという考えに基づいて，状態空間をタスク環境の状態と内部状態に分けて，それぞれの状態の価値を独立に推定することにしている．そして，タスク環境の状態の価値を状態クラスタの自動構築に利用して，各クラスタ毎に内部状態の価値推定と行動の決定を行わせるものとしている．本アルゴリズムの大まかな流れを図 1 に示す．以下では，独自のアルゴリズムである，状態クラスタの自動構築法と状態クラスタ内での行動決定 · 学習法について説明をしていく．

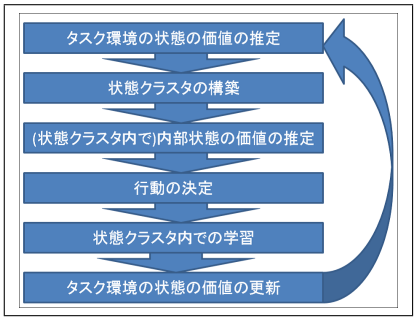

図 1: 提案アルゴリズム

2.1　状態クラスタの自動構築法

状態クラスタの構築は step 毎に行われ，そのタイミングは現在のタスク環境の状態の価値を推定した直後とする．状態クラスタ構築の際には，**拡張**，**追加**，**分離**，**統合**の処理がそれぞれ実行される．

拡張

現在の状態を 1step 前にエージェントがいた状態クラスタの要素にして，その状態クラスタの範囲を広げることである．この処理が実行される条件は，現在のタスク環境の状態の価値が 1step 前にエージェントがいた状態クラスタの「価値の信頼区間内にいる」ことである．「価値の信頼区間内にいる」とは，状態クラスタ内の状態の価値の分布を正規分布と仮定した時に，対象の状態の価値がその正規分布の信頼区間内に位置することを意味している．また，対象の状態の価値が正規分布の信頼区間から外れている場合は，「価値の信頼区間外にいる」と呼ぶことにする．

追加

現在の状態を要素とする状態クラスタを新たに作成することである．現在の状態がどの状態クラスタにも属さず且つ拡張が行われない場合，この処理が実行される．

Reinforcement Learning Using State Clusters in Continuous State Space.
†Michihiro Onodera・Graduate School of Tokyo University of Science.
‡Teruhiko Suzuki・Tokyo University of Science.
‡Ikuo Tahara・Tokyo University of Science.

分離

現在の状態がある状態クラスタに属している場合，現在の状態をその状態クラスタから切り離して，現在の状態を要素とする状態クラスタを新たに作ることである．現在の状態の価値が対象の状態クラスタの「価値の信頼区間外にいる」場合，この処理が実行される．

統合

現在と 1step 前の状態クラスタが異なる場合，両方のクラスタを 1 つのクラスタにまとめることである．クラスタ同士が近接していて且つ両方の状態クラスタの平均価値がお互いのクラスタの「価値の信頼区間内にいる」場合，この処理が実行される．なお，統合後のクラスタの学習器として，より学習回数の多いクラスタの学習器を使用していく．

2.2 状態クラスタ内の学習法

各状態クラスタの学習器には，連続状態行動空間に有効である actor-critic 法という学習アルゴリズムを採用した．actor-critic 法とは，状態の価値を評価する critic と状態から行動への確率分布である確率的政策に従って行動選択を行う actor の 2 要素から構成されている手法である [2]．本手法では，critic で内部状態の価値の推定を行い，actor に基づいてクラスタ内での行動を決定する．そして，critic において行動後に得られた報酬を元にして TD 誤差を求めて，その TD 誤差を用いて critic 部の状態の価値と actor 部の確率的政策を更新する．

しかし，この学習法のみでは，ゴール状態を要素として持つクラスタしか報酬を得られないので，そのクラスタでゴール状態へ近づく行動系列を獲得できたとしても，その他のクラスタではそのような行動系列を得ることはできない．そこで，ゴール時に各クラスタ A に次のような報酬 r_A を与えることにした．

$$r_A = r \times \gamma^{(t_G - t_A)} \tag{1}$$

ここで，r はゴール時に得られる報酬，γ は割引率 (値域は $0 \leq \gamma \leq 1$)，t_G はゴールに到達した時刻，t_A はエージェントが最後にクラスタ A 中の状態でいた時刻をそれぞれ表していである．この報酬を各クラスタに与えることによって，各クラスタで行われた行動に対してゴール到達までの貢献度に見合った報酬が与えられるので，各クラスタに適した行動系列を獲得することが可能になる．

3 実験

提案手法の有効性を検証するために，障害物が存在する目的地到達タスクを，通常の actor-crit 法を実装したエージェントと提案手法を実装したエージェントに行わせて比較した．状態空間として用いる状態は，エージェントの位置，視界，速度，ゴールとの角度の 4 種類である．エージェントの行動は，速度の変更と向いている角度の変更の 2 種類である．提案手法では，エージェントの位置のみを利用して状態クラスタを生成している．そして，両方のエージェントに対してこのタスクを 200episode 行わせた．

この実験の結果，actor-critic 法を実装したエージェントは 4episode までは目的地に到達できたものの，それ以降の Episode では障害物にぶつかったまま動けず目的地に到達できなくなった．一方，提案手法を実装したエージェントは全ての episode において目的地に到達することができた．しかし，提案手法を用いても学習が収束する (各 episode における目的地到達までの消費 step 数が一定値で落ち着く) ことはなかった．その原因としては，状態クラスタの自動構築の際に分離が何度も行われてしまっていることが考えられる．新しく生成される状態クラスタの学習器のパラメーターの値は初期値であるため，学習のために step を消費する必要が生まれて状態クラスタの数が増える程消費 step 数も増えてしまうからである．

4 おわりに

実験により，本提案手法を用いることによって連続状態空間上のタスクを遂行できることは検証できた．しかし，同時に学習が収束しないという問題点も浮き彫りになった．今後はこの問題点の解決に努めたい．

また，この提案手法によって生成された状態クラスタを内部状態の異なるエージェントに効率的に再利用させることも可能だと考えられるので，そうした応用法についても検討していきたい．

参考文献

[1] J.H.Connel，S.Mahadevan， Robot Learning，Kluwer Academic Publishers，1993.

[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.10，1999.