3B2-R-2-2

連続値入力問題のためのガウス型状態表現を用いたTD学習法

A Temporal-Difference Learning Method Using Gaussian State Representation for Continuous State Space Problems

藤井 菜摘子 Natsuko Fujii　上野 敦志 Atsushi Ueno　田窪 朋仁 Tomohito Takubo　辰巳 昭治 Shoji Tatsumi

大阪市立大学 大学院工学研究科 電子情報系専攻
Graduate School of Engineering, Osaka City University

When applying Reinforcement Learning to environments with continuous state spaces, a relevant discretization is required for avoiding perceptual aliasing. A method that uses Gaussian state representation for corresponding the problem is known. In this paper, we propose a method that applys the Gaussian state representation to TD learning for environments with continuous state spaces and noisy actions. We show the effectiveness of our proposal by computer simulations of a path finding problem.

1. 概要

強化学習で連続値の状態空間を持つ環境を扱う場合，連続値で与えられる知覚入力を離散化する必要がある．これを問題環境に適した形で行わなければ，同一視すべきでない複数の知覚入力を同じものとして認識してしまう不完全知覚問題が発生しやすくなり，学習が進まなくなってしまう．しかし，これをユーザが手動で設定するのは困難である為，状態空間の離散化を学習エージェントが学習過程で自律的に行う手法が提案されている．合理的政策形成アルゴリズム (RPM) を連続値入力に適用した手法 (連続値入力 RPM と呼ぶことにする)[1] ではガウス型の**状態領域**[*1] を用いて連続空間を離散化する．この手法は PS 法の特徴を生かし，素早く合理的な政策を得ることができるが，局所解に陥ってもより最適な政策を探索しようとはせず，さらに行動にノイズを多く含む環境ではうまく働かないという問題を持つ．また，連続値入力 RPM を罰情報に対応させた手法として罰回避政策形成アルゴリズム (PARP) を連続値入力に適用した手法 [2] が提案されているが，ゴール指向の問題に対応した手法ではない．そこで本研究では，この連続空間の離散化手法を DP 法の一種である TD 学習に適用した手法を提案し，本手法の有効性をコンピュータシミュレーション上の経路探索問題で示す．

2. 連続値入力 RPM

2.1 状態領域による連続値入力の取り扱い

連続値入力 RPM では，時刻 t の知覚入力を $\boldsymbol{s_t}$，$\boldsymbol{s_t}$ で選択した行動を a_t，その結果遷移した先を $\boldsymbol{s_{t+1}}$ とすると，$\boldsymbol{s_t}$ から $\boldsymbol{s_{t+1}}$ に遷移した時点で，図 1 に示すような $\boldsymbol{s_t}$ を中心とする n 次元正規分布関数により状態領域を生成する．n は状態空間の次元数である．領域の主軸 (図 1 の d_1 軸) は移動方向で，主軸以外の方向 (図 1 の d_2,$\cdots$,d_n 軸) は，各々が直交するようにグラムシュミットの直交化法を用いて生成する．主軸の裾野の広さは，$3\sigma_1 = |\boldsymbol{s_{t+1}} - \boldsymbol{s_t}|$，主軸以外の裾野の広さは，$3\sigma_i = \frac{|\boldsymbol{s_{t+1}} - \boldsymbol{s_t}|}{\sqrt{n}}$(i=2,3,$\cdots$,n) とし，移動方向に引き延ばした形を実現する．状態領域を生成する際，選択した行動 a_t を各状態領域に記録し，行動選択時に利用する．

連絡先: 藤井 菜摘子，大阪市立大学大学院工学研究科電子情報系専攻，fuji@kdel.info.eng.osaka-cu.ac.jp

*1 元論文の**基底関数**と同意．

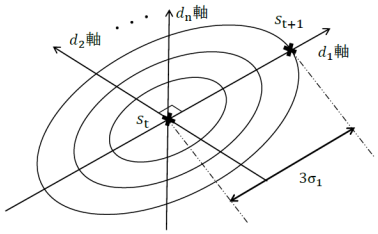

図 1: 状態領域の形状 (文献 [1] の図 1 を参考に作成)

最初の報酬を得るまでは一様ランダムな確率で行動を選択し，状態領域を生成していく．ここで，得られた知覚入力といずれかの状態領域の中心 $\boldsymbol{\mu}$ とのユークリッド距離が定数 l_{near} 以下となった場合，既存の状態領域に上書きする．報酬を得た後は，ある知覚入力 $\boldsymbol{d}$ が観測されたとき，どの状態領域に属するかを以下の評価式で調べる．

$$f(\boldsymbol{d}) = e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{(\mu_i - d_i)^2}{\sigma_i^2}} \quad (1)$$

$\boldsymbol{d}$ が $\boldsymbol{\mu}$ に一致したとき $f(\boldsymbol{d})$ は 1.0 となり，$\boldsymbol{d}$ が $\boldsymbol{\mu}$ から離れるほど小さな値をとる．$f(\boldsymbol{d})$ が各状態領域に設定されたしきい値 f_{para} 以上で最も大きくなる状態領域が知覚入力 $\boldsymbol{d}$ に対応する離散状態とし，各状態領域に記憶された行動を選択する．$f(\boldsymbol{d})$ がしきい値より大きな状態領域がなければ，一様ランダムに行動を選択し新たに状態領域を生成する．

ある行動数経過しても報酬が得られないならば，その政策は合理性を維持していないと判断し，その報酬の得られないループに含まれる状態領域の f_{para} を大きくする．この処理により領域の守備範囲が狭まるため，結果として不完全知覚が減少する．f_{para} の更新の結果，$f_{para} = 1.0$ となる状態領域が含まれた場合，不適切な領域が生成されている可能性があるので，すべてを初期化するマルチスタート法 [3] を採用し，学習を最初からやり直す．

2.2 連続値入力 RPM の問題点

ノイズを含む問題環境では行動の結果が分散するので，通常なら前進できない行動 (失敗行動) でも稀にうまくいくことがありえる．その場合にループ系列が形成されやすく，従来手法ではエピソードが失敗と判定されるまで失敗行動を繰り返

し，学習に不要な状態領域を多数生成してしまう．その為，エピソード内でループから抜け出す手段が必要となる．また，従来手法では不必要な状態領域に関して f_{para} を大きくすることで徐々に対応する状態領域を小さくしていくため，状態領域の生成能力に比べて，不要な状態領域の削除能力が弱い．

3. 提案手法

2.2 節で述べた問題は，各状態領域に価値を持たせ，TD 学習によって価値計算を行うことで対応可能である．提案手法では状態領域それぞれに価値を与え，TD 学習で価値を更新する．価値の更新は以下の TD 学習の更新式に従う．

$$V(\boldsymbol{s_t}) \leftarrow V(\boldsymbol{s_t}) + \alpha[r_{t+1} + \gamma V(\boldsymbol{s_{t+1}}) - V(\boldsymbol{s_t})] \quad (2)$$

α は学習率，γ は割引率，r は報酬である．報酬はゴール時の成功報酬と衝突およびループの罰の三種類とする．行動後の知覚入力に対応する状態領域がない場合には $V(\boldsymbol{s_{t+1}}) = V_{init}$ とする．V_{init} は価値の初期値である．

知覚入力 $\boldsymbol{d}$ が与えられたとき，式 (3) により各状態領域の得点 p を求め，得点 p が定数 p_{border} 以上であれば対応しているとし，対応する状態領域の中で得点 p が最も大きい状態領域が，現在の知覚入力に最も近い状態領域であると判定する．

$$p = f(\boldsymbol{d}) \cdot V(\boldsymbol{s}) \quad (3)$$

p_{border} は正の定数であるため，価値が負となった状態領域は自動的に選択されなくなり，非アクティブとなる．また，状態領域の使用回数をエピソードごとに記憶しておき，各ステップごとに，エピソード内の状態領域の使用回数が定数 N_{use} 以上になった場合に使用回数が $N_{use} - 1$ 回以上の全ての状態領域をループ系列であると判定して負の報酬を与える．これによってループ系列から抜け出すことができる．

4. 実験

4.1 実験設定

図 2 および図 3 に示した 2 つの 2 次元平面環境において実験を行う．始点の座標から行動を開始し終点に到達すると報酬が与えられ始点に戻されるものとする．感覚入力は，x および y 座標の 2 次元連続値で与えられ，行動は上下左右に斜め方向を加えた 8 種類から選ばれる．各方向への移動量は 0.1 の基本量にランダムな方向に 0〜10 %のノイズが加わる．外枠および灰色部分には進入不可であり，進入しようとした場合は元の位置に戻される．

提案手法との比較のために，連続値入力 RPM および格子空間を用いた Q-learning でも性能評価を行う．Q-learning では x 軸および y 軸方向をそれぞれ 10 個および 20 個に区切り，合計 100 個または 400 個の格子で状態空間を離散化する．提案手法と Q-learning では終点に到達した際に 1.0 の正の報酬，壁に進入しようとした際に-0.1 の負の報酬を与える．また，提案手法ではループ系列に陥ったと判定された際に-0.5 の負の報酬を与える．迷路 1 の最大行動数は 1000 回，迷路 2 の最大行動数は 3000 回とする．100 回試行を行い，結果の平均をとる．実験で用いたパラメータは以下の通りである．提案手法に関しては，$\alpha = 0.5$，$\gamma = 1.0$，$V_{init} = 0.1$，$p_{border} = 0.001$，$N_{use} = 3$，$l_{near} = 0.05$，連続値入力 RPM に関しては $l_{near} = 0.0005$，Q-learning に関しては $\alpha = 0.5$，$\gamma = 0.9$，$V_{init} = 0.1$，$\epsilon = 0.1$ とした．

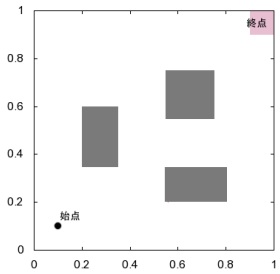

図 2: 迷路 1

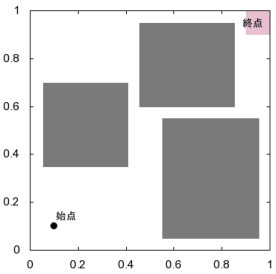

図 3: 迷路 2

4.2 結果

図 4 にノイズなしの迷路 1 の結果を示す．提案手法は他手法よりも学習が早かったものの，収束ステップ数は 21.72 となり，Q-learning の 18.14 よりも大きくなった．これは Q-learning の入力空間が適切に離散化されていることと，提案手法が経験にこだわって十分な最適化が行われていないことを意味する．しかしその差は微小であり，提案手法でも十分に良い解が得られている．次に，図 5 にノイズありの迷路 1 の結果を示す．提案手法および Q-learning はランダム性のある環境でもうまく学習を行えたが，連続値入力 RPM はよい収束値を得ることができなかった．これは 2.2 節で述べたように，エピソードが失敗と判断されるまでループ系列を繰り返し続けるような状態領域を作成し続けているためであると考えられる．

図 6 にノイズなしの迷路 2 の結果を示す．離散化が荒すぎたために格子数 100 の Q-learning の結果が悪くなり，格子の数を 400 に増やすと学習が進んだものの，状態数が大きくなり学習の速度が提案手法より遅れた．最後に，図 7 にノイズありの迷路 2 の結果を示す．提案手法が全手法の中でもっとも良い性能を得た．連続値入力 RPM は迷路 1 のノイズありの問題と同様うまく学習を行えず，格子数 400 の Q-learning も学習速度が遅くなった．

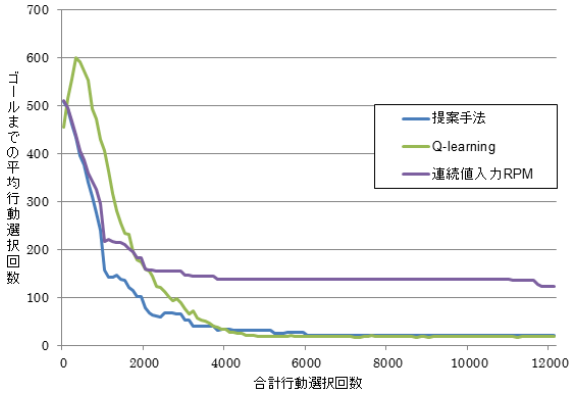

図 4: ノイズなし迷路 1 の結果

5. 考察

いくつかの実験で Q-learning に収束値が劣ったものの，提案手法は全ての実験を通して安定した良い性能を得られた．これは，連続値入力 RPM で提案されたガウス型の状態領域による離散化手法が，TD 学習においても有効に動作することを示

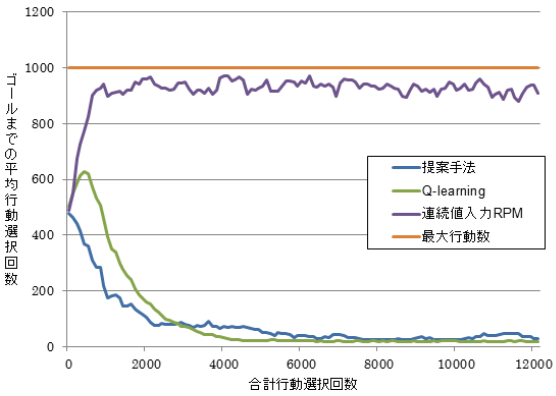

図 5: ノイズあり迷路 1 の結果

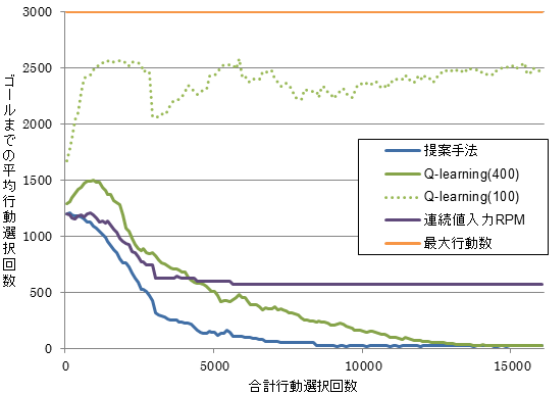

図 6: ノイズなし迷路 2 の結果

している．不完全知覚問題に頑強でないとされる TD 学習で学習が進んだのは，提案手法によって適切に離散化が行われ，結果不完全知覚の発生を抑制できたためであると考えられる．今回の実験で作成した状態領域の様子を図 8 に示す．迷路 1 のノイズあり 8 方向において，150 エピソード経過後に得られた状態領域の様子である．各状態領域に関して，得点 p が 0.3 以上となった領域部分を図示している．始点の周囲と終点の周囲において領域の大きさが等しいのは，TD 学習の更新式において割引率 γ を 1.0 としているため，割引が行われず収束値が等しくなるためである．学習で得られた経路以外にも状態領域が生成されているが，学習過程においてふさわしくないとされる経路上の状態領域は，価値の更新が途中で止まり状態領域が小さくなることが分かる．

6. 結論

連続値の状態空間を持つ問題環境に強化学習を適用する場合，状態を適切に離散化することが重要となる．本研究ではガウス型の状態表現を用いて学習エージェントが学習過程で自律的に離散化を行う手法と TD 学習を組み合わせることで，連続値の状態空間を適切に離散化できることを，実験を通して示した．

連続値入力 RPM では，ノイズなどの影響で行動の失敗を繰り返す問題や，不必要な状態領域を余分に生成し続けるなどの問題があり，十分に学習が行われなかった．一方提案手法ではステップごとに価値の更新を行うことで必要な状態領域の取

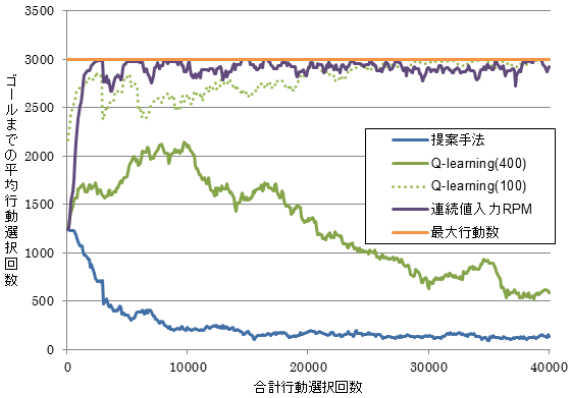

図 7: ノイズあり迷路 2 の結果

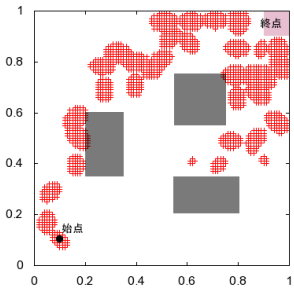

図 8: 作成した状態領域の様子

捨選択を行うことができ，行動出力にノイズを含む問題でも適切に学習を行うことができた．また，格子による離散化を用いた Q-learnig では格子の大きさを問題環境ごとに適切に決定する必要があったが，提案手法では学習エージェントが自律的に離散化を行うため，どのような問題環境にも対応することができた．また，提案手法は学習に必要な領域に関する情報のみで学習を行うことができ，格子空間による離散化と比べてより効率的に学習を行うことができる．この特徴は多次元の問題環境でさらに有効に動作すると考えられ，多次元問題への適用は今後の課題となっている．

参考文献

[1] 宮崎和光，木村元，小林重信 : 合理的政策形成アルゴリズムの連続値入力への拡張，人工知能学会論文誌，Vol.22, No.3，pp.332-341（2007）.

[2] Kazuteru Miyazaki and Shigenobu Kobayashi : Reinforcement Learning for Penalty Avoidance in Continuous State Spaces，Journal of Advanced Computational Intelligence and Intelligent Informatics，Vol.11，No.6，pp.668-676（2007）.

[3] 宮崎和光，荒井幸代，小林重信 : POMDPs 環境下での決定的政策の学習，人工知能学会誌，Vol.14，No.1，pp.148-156（1999）.

[4] 河合宏和，上野敦志，辰巳昭治 : POMDPs 環境のための報酬獲得効率に基づく強化学習法，人工知能学会論文誌，Vol.23-A，No.1，pp.1-12（2008）.