ISSN 2186-5647



P-23

# ノイズを含む連続値入力問題における最適性の獲得

日大生産工　○杉浦 慶久　日大生産工　山内 ゆかり

## 1 まえがき

強化学習では、報酬を得ることで状況に対応することが可能となっていく機械学習の一種である。この学習法は主に、迷路探索問題といった状態空間を扱う問題で活用されている。

この状態空間に関する問題には、連続値で与えられた入力空間を扱うものがあり、連続値入力問題と呼ばれる。

これらを扱うための手法として、格子表現や基底関数表現などのユニット群である状態表現を求め、その状態表現に対して重みを付けることにより、連続状態空間において適切な行動を求めることが可能となる[1]。この手法を用いると、適切に状態表現を設定した場合には良い学習性能を得ることができる反面、状態表現の数が不足しているか、逆に多過ぎると、正しく学習されないという問題がある。

そこで宮崎らは、経験強化型状態表現学習を行う手法では、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[2]。この連続値入力RPMと組合せて罰を得た場所を記憶することで、罰を得る行動を回避することができる。これを罰回避政策形成アルゴリズム(PARP)と呼ぶ。このPARPでは一度得た合理的政策よりも良い政策を探索しないため、ゴールまでの政策を素早く得ることができる。しかし、性質上、一度獲得した経路から外れた場所では政策を得られないため、ノイズによって誤差が生じる環境では正しい学習を行うことができない。

また、生成した状態のうち不要な状態の削減能力が弱いという問題がある。

そこで藤井らの連続値入力問題のためのガウス型状態表現を用いたTD学習法の研究では、それらの問題に対応するために、生成される状態に対して価値の概念を導入した[3]。この価値を不要な状態の素早い削減のために用いることにより、ゴールに到達した経路上の状態の価値は高くなり、逆に、ノイズによる誤差によって発生する、ループや壁への衝突などの不適切な状態には罰が与えられ、価値が下がる。価値の高いものは行動の際に選択されやすくなるが、低い場合は領域が徐々に削減されていく。これにより、行動にノイズが含まれる問題でも素早く合理的な政策を得ることが可能となった。

しかし、連続値入力RPMおよび従来研究では、一度得た合理的政策以上の政策を探索しない関係上、解の最適性が保証されないという問題点があった。

本研究では従来研究の合理的政策に加え、最適な解を得ることを目的とする。具体的には、複数の学習器による協調行動を提案する。

継続的に学習を行う主学習器1つと、探索を目的として行動する副学習器複数に分け、同一環境中で行動および学習を行う。主学習器が一定回数学習を終えた後、副学習器が未知環境として探索を行い、その結果得た優れた情報を主学習器と共有していく。これを繰り返すことで、最終的に最適性を得ることができると考える。

## 2 従来研究

2.1 藤井らは状態表現を学習で獲得する手法の省力性と効率性に着目し、連続値入力RPMに対して状態価値の学習法であるTD学習法を追加した[3]。価値$V(s_t)$の導出式を式(1)に示す。

$$V(s_t) \leftarrow V(s_t) + \alpha[\, r_{t+1} + \gamma V(s_{t+1}) - V(s_t)\,] \tag{1}$$

次にRPMのアルゴリズムを示す。

1. 1次および2次を合わせた合計$2N$個分の記憶領域を初期化する。
2. 各時間の感覚入力の2次記憶上に行動が記憶されていればその行動を出力し、そうでなければ環境探索戦略による行動を出力し、その行動を1次記憶上に出力する。
3. 報酬を得た場合、その時点までの1次記憶領域の内容を全て2次記憶領域に複写する。
4. 2次記憶領域が収束し、合理的政策が得られている場合、その内容を保存する。

合理的政策が得られたら、それ以降は2次記憶に従い特定の行動を選択するので、それ以降の探索は行われない。

---

Optimality of the Acquisition in Continuous Value Input Issues, Including Noise

Yoshihisa SUGIURA and Yukari YAMAUCHI

2.2 連続値入力問題におけるRPMの利用

学習器が意志決定を行うごとに、基底関数(状態)を生成し、$n$次元の連続値で与えられる感覚入力を離散化する。

時刻$t$における感覚入力を$s_t$、$s_t$で選択した行動を$a_t$、行動により遷移した先を$s_{t+1}$とする。遷移した時点で、$s_t$を中心としたn次元正規分布関数により基底関数を生成する。

関数の主軸方向は$s_{t+1}-s_t$とし、それ以外の方向は各々が直交するように生成する。

主軸の裾野の広さは$3\sigma 1=|s_{t+1}-s_t|$、それ以外は$3\sigma 1=|s_{t+1}-s_t|/\sqrt{n}(i=2,3,\dots,n)$とする。

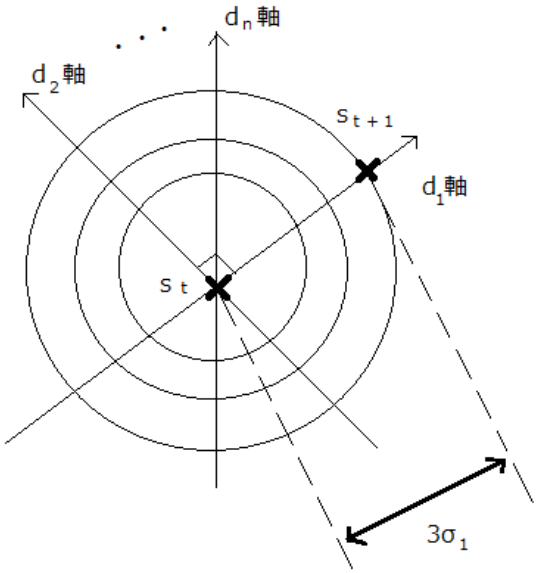


**図1.基底関数の形状**

・関数の返り値の最大値$f(d)$

最大値が1.0になるように変換する。$f(d)$の導出式を式(2)に示す。

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

$d$が$\mu$と一致したとき1.0となり、離れていくほど小さな値となる。

・状態の生成と行動選択

最初の報酬を得るまでは一様ランダムに行動を選択し、状態を生成する。ここで、ある時点での入力と、既存の基底関数の状態$\mu$とのユークリッド距離が、ある非常に小さな値以下となった場合には、既存の関数に上書きする形で新たな関数を生成する。

入力を得た後は、その入力が既知のいずれの基底関数に属するかを判定する。判定には$f_para$を参照し、各基底関数から返される(2)式の値が各関数の$f_para$以上であれば、その関数は既存の感覚入力に近い基底関数とされる。

行動は、近い基底関数の中で(2)式の値が最大の関数を選択する。近い状態がなければ、一様ランダムに行動選択を行う。その場合、新たな基底関数を生成する。

・連続値入力RPMにおけるループ対策

一定数行動してもゴール領域に到達しなければ、その政策は不適切と判断し、その行動経路上の状態の$f_para$を大きくする。これにより、不適切な行動選択を行う状態の守備範囲が狭まる。ある状態で$f_para$の値が1.0以上となった場合、不適切な状態が生成されている可能性が高いので、マルチスタート法によって学習状況を初期化する。

2.3 ガウス型状態表現を用いたTD学習法

ノイズを含む連続値入力環境に対応するため、従来研究では連続値入力RPMに状態価値の概念を導入した。この状態の価値は不要な状態を素早く削除するために使用される。

従来研究では各状態に価値Vを持たせ、行動選択を行う際に利用する。その状態の導出式を式(3)に示す。

$$s:=<\mu,\Sigma,\alpha,\sigma^2,V> \qquad (3)$$

式(3)について、各状態$s$は中心座標$\mu=(\mu_1,\dots,\mu_n)^T,d_1,d_2,\dots,d_n$軸空間への座標変換行列$\Sigma$、行動$\alpha$、分散$\sigma^2=(\sigma_1^2,\dots,\sigma_n^2)^T$、価値$V$で構成される。$n$は入力空間の次元数、価値$V$の初期値は定数$V_{init}$で与えられる。

・価値更新

一時記憶の全状態において、TD学習を用いて価値更新を行う。状態$s_t$から状態$s_{t+1}$に遷移し報酬$r_{t+1}$を受け取った時の更新式は式(1)で表される。

状態$\alpha$は学習率、$\gamma$は割引率である。行動後の入力に対応する状態がない場合には$V(s_{t+1})=V_{init}$とする。

価値の更新によって、ゴールに到達した経路上の状態の価値は高くなり、選択の際に選ばれやすくなる。逆にノイズの影響で偶然得た状態は価値が高くならず、ループに陥るような状態

や、壁に衝突する状態には罰が与えられ価値が下がるため、領域が徐々に削減されていく。

・状態の選択
入力$d$が与えられたとき、主記憶内の各状態sにおいて、式(2)により与えられる$f(d)$と価値$V(s)$から、以下に示す式(4)により得点$p$を求める。

$$p = f(d) \cdot V(s) \tag{4}$$

この$p$が定数$p_{border}$以上かつ最大の状態が、現在の入力に対応する状態であると判定する。

・従来研究におけるループ対策
連続値入力RPMのアルゴリズムでは、ノイズを含む環境では不適切な状態が生成されることでループに陥る可能性が高くなる。対策として、各状態にエピソード内での行動選択回数を記憶し、ある状態で選択回数が定数$N_{use}$以上になった場合に選択回数が$N_{use}-1$回以上の全ての状態に対して負の価値を与える。これによって、何度もループに陥る状態があれば、価値更新によって価値が負になり淘汰される。

## 3 提案手法

本研究では、従来研究の最適性が得られないという特徴を解決するべく、従来研究に対して複数の学習器による協調行動を追加することを提案する。

学習器は、継続的に学習を行う主学習器１つと、探索を目的として行動する副学習器複数に分けられる。これらを用いて学習を行う。下記に提案手法のアルゴリズムを示す。

1. 主学習器を環境中で行動させ、従来手法同様のアルゴリズムを実行する。
2. 主学習器がゴール領域に到達し、報酬を得た場合、その時点で一旦主学習器を停止し、代わりに副学習器を同一環境中で探索させる。これを用意した副学習器全てが1エピソード分を終了するまで実行する。
3. 副学習器が得た結果と主学習器の結果を比較し、副学習器の内いずれかが主学習器よりも短いステップ数でゴール領域に到達した場合、その副学習器と主学習器に記憶されている各座標情報を探索し、一致する座標の主学習器の状態における価値と行動を副学習器のものに更新する。一致しない座標の状態が存在する場合、主学習器に対して、その状態を追加する。
4. 1～3を繰り返し実行する。

## 4 実験環境

本実験では、以下の2次元平面環境の迷路探索問題を用いて実験を行う。

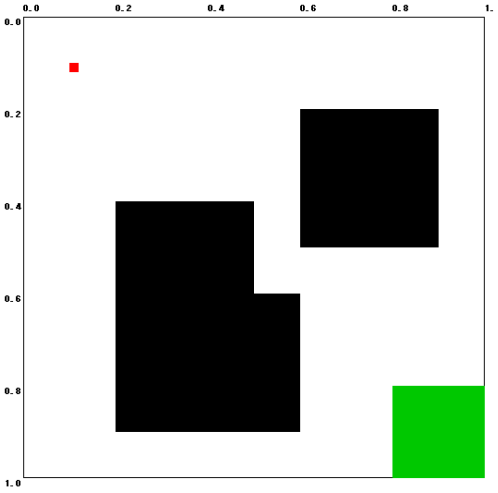


**図2.実験環境**

スタート地点($x$=0.1,$y$=0.1)より行動を開始し、ゴール領域(右下の四角領域、本実験環境では$0.8 \leqq x \leqq 1.0$かつ$0.8 \leqq y \leqq 1.0$)まで到達すると正の価値を得て再びスタート地点まで座標がリセットされる。この一連の流れを1エピソードとする。

この過程に至るまでの平均ステップ(行動)数を短くすることを目標とする。なお、壁に衝突する行動に対しては負の価値が与えられる。進行方向は上下左右に斜め方向を加えた8方向であり、さらに進行方向に加えてランダムに0～10%のノイズがランダム方向に発生する。
最大ステップ数は3000回までとし、それ以上はエピソードを失敗と判断してスタート地点へ戻される。

また、連続値入力RPM、従来研究を同様の環境で実験し、提案手法との比較を行った。各手法をエピソード10回分実行した結果の平均値を以下に示す。ただし、連続値入力RPMには価値の概念が存在しないため、一連の実験に価値は影響しない。

提案手法については、主学習器のみのエピソードを実験に用いるものとする。なお、この提案手法は項目3の比較および共有を実装していない他、項目2についても正常な動作が確認出来ていないため、現在の状態で実験を行った結果を記載する。

以下、各手法に共通する、数エピソード経過時の実験例を図3に示し、表1に各手法の実験結果を示す。

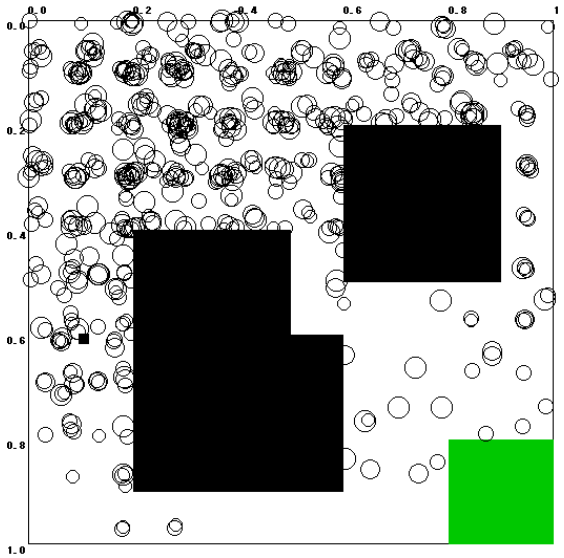


**図3. エピソード経過時の状態生成例**

学習器が行動を選択しながら、状態を生成していく。生成された状態は、次のエピソードを実行する際に利用され、また、状態が存在しない位置には再び新たな状態が生成され、さらに次のエピソードを実行する際に利用される。この繰り返しによって連続値入力問題に対応する学習が可能となる。

**表 1. 各手法における平均行動回数**

| エピソード | 1回目 | 3回目 | 10回目 |
|---|---|---|---|
| 提案手法 | 489.67 | 449.33 | 236.67 |
| 従来研究 | 526.67 | 191.33 | 41.00 |
| 連続値入力 RPM | 676.67 | 51.00 | 56.00 |

表 1 の結果について説明すると、1 回目は一様ランダムに行動してゴール領域に到達した際の平均ステップ数、2 回目以降は 1 回目の学習内容に従って行動した際の平均ステップ数となっている。学習を繰り返す内に、次第に結果は収束していく。

連続値入力 RPM は比較的安定した結果を得られているが、これはノイズの影響で偶発的に適切な解を得たに過ぎず、あるエピソード中に最大ステップ数までゴール領域に到達することができないことがあった。

従来研究はノイズを含む環境においても問題なく学習でき、安定した解に収束している。しかし、問題点として挙げた通り、今回の結果では最適性を得られていない。本実験環境では、20 ステップ以内にゴール領域に到達することが可能なためである。

そして提案手法は、現時点では正しく動作していないため、上述の他手法から大きく離れた結果となった。

## 5 まとめ

本研究では、強いノイズを含む連続値入力環境に対して、従来研究の問題点のひとつであった最適な経路を獲得することを目的とした複数の学習器による学習を提案した。

正常に動作した場合、主学習器は探索によって得たステップ数がより短い副学習器の情報を共有し、主学習器自体も学習を行うことにより、最終的に最適な解を得ることができると考える。

提案手法の今後の課題は、副学習器の探索と、それによって得られた情報の共有を主学習器と行う部分を実装することである。

「参考文献」

1) 宮崎和光, 木村元, 小林重信, 合理的政策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 AI 22, pp.332-341, 2007-11-01
2) 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 14(1), pp.148-156, 1999-01-01
3) 藤井菜摘子, 上野敦志, 田窪朋仁, 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, 人工知能学会論文誌 29(1), pp.157-167, 2014