ISSN 2186-5647

P-8

ベータ分布伝搬型強化学習の探索性能の向上

日大生産工　○古賀　博也　　日大生産工　山内　ゆかり

1. まえがき

強化学習は環境との試行錯誤を通じて状態に対する正しい行動を得るための学習制御の枠組である。環境から望ましい状態に対して報酬が与えられることで、学習エージェントはそれに関連する状態、行動を評価し、報酬獲得までの到達プロセスを得る。設計者が最終的にさせたいことを報酬という形で指示しておけば、そのためのプロセスを学習主体が自動的に獲得する枠組みとなっている。

野津らは、行動政策と価値推定をベイズ推定から捉え直し、ベータ分布を状態行動対の価値として伝搬させ強化学習させる手法（Beta Distribution Propagation Learning：BDPL）を提案した。1)しかし、探索が少ない状態で活用の段階に入ってしまうなど状態遷移が不安定である問題がある。

そこで本研究では、BDPLのアルゴリズムを改良し、探索状態を数値化及び状態遷移の最適化を行うことで、探索性能を向上させることを試みる。

2. 従来手法

2.1. ベータ分布伝搬による学習

プロスペクト理論から得られる知見を応用し，ベータ分布を状態行動対の価値として伝搬させ強化学習させる手法（Beta Distribution Propagation Learning：BDPL）はトンプソンサンプリング、ベータ分布によるベイズ推定をプロスペクト理論に基づいて強化学習として構築したものになる。累積の正の報酬をベータ分布のパラメータα、累積の負の報酬をパラメータβと考えることで各状態行動対の価値をベータ分布で表現する。

2.2. 良い経験の伝搬

ある状態である行動をしたときに、トータルでどれくらい良いことがあったのかを示す数値をαとする。現状態s、次状態s'、行動a、正の報酬r_+ のとき、パラメータ$\alpha(s,a)$の更新式を以下のように定義する。

$$\alpha(s,a) = \gamma_{\alpha/\beta}\alpha(\gamma_{now}\alpha(s,a) + \gamma_{next}\alpha(s',a') + r_+) \quad (1)$$

$$a' = \underset{a^*}{argmax}\ \frac{\alpha(s',a^*)}{\alpha(s',a^*)+\beta(s',a^*)} \quad (2)$$

次状態で期待値が最大となる選択肢a'の$\alpha(s',a')$と現状態行動対の$\alpha(s,a)$をそれぞれγ_{next},γ_{now}を乗じて減衰させ和をとったものに正の報酬r_+を加え、良い経験の重み付き累積和とする。

γ_{next}は次状態の価値をどの程度伝搬させるかを示すパラメータであり、γ_{now}は現状態をどの程度覚えているか、というパラメータになる。

2.3. 悪い経験の伝搬

ある状態である行動をしたときに、トータルでどれくらい悪いことがあったのかを示す数値をβとする。$\beta(s,a)$の値についてもαと同様に、負の報酬の大きさがr_-のとき

$$\beta(s,a) = \gamma_{now}\beta(s,a) + \gamma_{next}\beta(s',a') + r_- \quad (3)$$

とする。

2.4. 行動選択

状態sで以下の式$Value(s,a)$が最も大きくなる選択肢、行動aを選択する。

$$Value(s,a) = E(s,a) + C\sqrt{V(s,a)} \quad (4)$$

$$E(s,a) = \frac{\alpha(s,a)}{\alpha(s,a)+\beta(s,a)} \quad (5)$$

$$V(s,a) = \frac{\alpha(s,a)\beta(s,a)}{(\alpha(s,a)+\beta(s,a))^2(\alpha(s,a)+\beta(s,a)+1)} \quad (6)$$

Cは分布のばらつきをどの程度考慮するかを調節するパラメータである。Cの値が0であればgreedyに期待値の高い選択肢を選択する手法となる一方、Cの値が極端に大きいと分散の大きい選択肢を選択するようになり、結果として全ての選択肢をほぼ等回数選択することになる。

3. 提案手法

本研究では、$Value$値が最大の状態を保存し、規定サイクルに到達したのち、最大の値を代入することにより、報酬値が高いものを選択するようにする。それにより報酬値が高いものにエージェントが移動することが期待できる。

Improved search performance of Beta Distribution Propagating Reinforcement Learning Based

Hiroya KOGA and Yukari YAMAUCHI

4. 実験方法および測定方法

本実験では障害なしのマップと迷路状のマップを作成し、離散空間最短経路探索を行う。

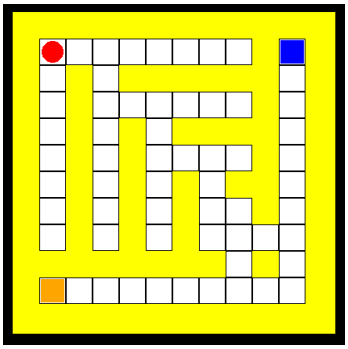

Fig.1 迷路状のマップ

実験に使用するマップは 10×10 障害のある迷路状のマップ(Fig.1)ある。

エージェントはスタート位置S_0から行動を始める。エージェントは各状態で上下左右いずれかの行動を選択し、ゴールに達すると報酬がもらえる。ゴールに到着または規定回数 100 回移動後にサイクルが終了し、エージェントはスタート地点に戻る。Fig.2 で、ゴールが二つ (G1、G2) で片方から$R_1 = 10$、もう一方から$R_2 = 1$の報酬を与えられる迷路状のマップの経路探索問題について比較する。これらを比較することで最短プロセスの学習だけでなく、より高い報酬を求めて効率よく状態空間を探索できているかを確認することができる。

本実験では200サイクル行い、最短行動回数、探索の広さ、柔軟性の確認をする。

5. 実験結果および検討

今回の実験で検討すべきことは二つである。一つはそれぞれの平均行動回数。もう一つは平均報酬率である。平均行動回数をFig.2、平均報酬率をFig.3に示す。

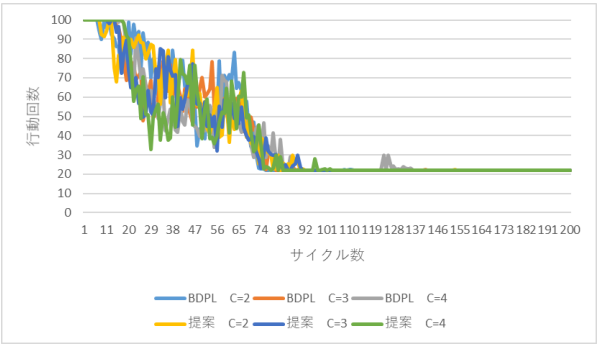

Fig.2 平均行動回数

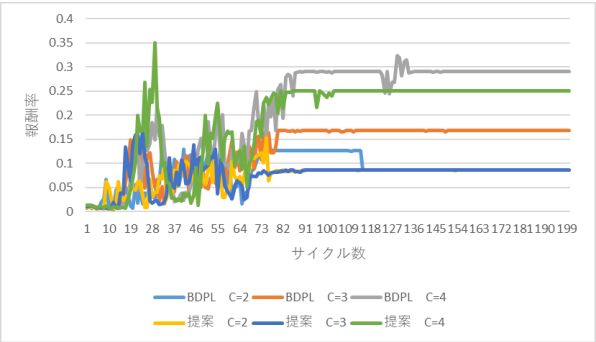

Fig.3 平均報酬率

従来手法であるBDPLと提案手法を比べてみる。まず、Fig.2の平均行動回数を見てみると従来のBDPLと提案手法を比べてみると行動回数は従来と同様の動きを見せている。これは規定サイクル到達後、同じ報酬値で行動していると見られる。

Fig.3の平均報酬率を見てみるとBDPLと比べて平均報酬率の向上は見られず下がっている傾向がある。

6. まとめ

規定回数に到達後に最大の値を代入することにより精度の向上が見られなかった、これは規定回数後の$Value$値の更新するとき値がすでに大きいからと考えられる。

今後の課題としては、過去の経験報酬の最大値を利用して、報酬を相対的に評価する仕組みを導入し、探索の段階で報酬値の高いルートのみの評価を上げる方法を検討していく必要がある。

参考文献

1) 野津亮，生方誠希，本田亮宏，プロスペクト理論を応用したベータ分布伝搬型強化学習による効率的探索と活用，知能と情報(日本知能情報ファジィ学会誌)Vol.29，No.1，2017，pp.507-516