ISSN 2186-5647

P-4

伝搬報酬(Propagated Reward)を利用した PRDUCBQ-learning

日大生産工(学部)　○楠田 正隼　日大生産工　山内 ゆかり

1 まえがき

近年の人工知能を利用した技術は,日々進歩し続けている.第三次産業から始まり,第二,第一次産業へも多用され始めている.第二次産業の代表として自動車への応用が期待されている.例えば,経路生成からの自立走行や画像認識によるオートブレーキ機能など人工知能が自ら判断し自動運転機能を持つ自動車も作られた.人工知能は今や身近な存在となりつつあり,研究が進められている.その一つとして野津らが提案した「Discounted UCB1-tunedのQ学習への適用」がある.これはUCB(Upper Condence Bound)AlogorithmをQ学習に応用した場合,変化する環境にどのようにバランスを取るかを問題とした.この問題に対して,Discounted UCBによる価値推定値の重み付き平均を用いることで変化する環境に対して学習を可能とした.しかし,人工知能において度々出てくる問題として局所解に陥る現象は野津らの論文においても例外ではない.

本研究ではモテンカルロ法を利用したMC(Monte Carlo)DUCBQ-learningを提案し野津らが提案した手法における局所解を避ける方法を検討する.

2 研究背景および従来手法

野津らは「Discounted UCB1-tunedのQ学習への適用」を提案し,UCBアルゴリズムとQ学習を組み合わせた手法の環境対応による改善を試みている.1)

野津らが提案したDiscounted UCBQ Alogorithmを示す.

1. $InitializeQ, V, UCB, N_a, NN_a$
2. 3→14 までを MAXCYCLE 回分繰り返す
3. $s = s_0$
4. 5→14 までを繰り返し,Time 回分繰り返す
   状態sとゴールs_gが一致すると抜ける
5. N_aとNN_aを条件にした if 文
   →満たす場合,状態更新したことない状態の行動をランダムに選択する
   →満たさない場合,maxarg$UCB(s,a)$を選択する
6. $NN_a(s,a) = NN_a(s,a) + 1$
7. 報酬を獲得する
8. 次状態s'へ遷移する
9. $s \neq s'$ならば 10 に,他は 14 に遷移する
10. $N_a(s,a) = N_a(s,a) + 1$
11. $Q(s,a)$値更新する
12. 分散値$V(s,a)$を更新する
13. 状態更新回数$N_a(s,a)$が一回以上ならば$UCB(s,a)$を更新し,それ以外は 0 とする
14. $s = s'$

Step1ではQ値$Q(s,a)$,分散$V(s,a)$,UCB値$UCB(s,a)$,状態更新回数N_a,行動選択回数NN_aが初期化される.また,初期エージェントの状態はランダムな初期座標が設けられ,相対的なゴール座標の状態s_gが設定され,現在のエージェントの状態sはs_0となり,状態sと状態s_gが直径D(=0.01)以下の時がゴールとなる.

Step4ではゴールでなれけば,以下の処理を行う.ゴールであればMAXCYCLE数を増やし,ゴールがランダムな座標へ再設定される.

Step5ではN_aとNN_aの条件を満たす場合は状態更新回数N_aが0の行動aを取り,これを探索段階とする.満たさない場合は$UCB(s,a)$における最適方策によって行動aを決め,これを学習段階とする.その後取った行動aによる報酬を得て,次状態s'に遷移を行う.状態遷移がおこらなかった場合はStep14に移る.

Step11では$Q(s,a)$を式(1)で更新を行う.

$$Q'(s,a) = Q(s.a) + \alpha[R_t + \gamma max_{a' \in A(s')} Q(s',a') - Q(s,a)] \quad (1)$$

Step12では分散値$V(s,a)$を式(2)で更新を行う.

UCBQ-learning by Propagated Reward

Masato KUSUDA　and　Yukari YAMAUCHI

$$V(s,a) = \frac{(\frac{r\left(1-r^{N_a(s,a)-1}\right)}{1-r})(V(s,a)+Q(s,a)^2}{\frac{1-r^{N_a(s,a)}}{1-r}} + \frac{(R_t+\gamma max_{a'\in A(s')}Q(s',a')^2}{\frac{1-r^{N_a(s,a)}}{1-r}} - Q'(s,a)^2 \quad (2)$$

その後状態sでの全ての行動aに対して,N_aの条件を満たす場合,Step13では$UCB(s,a)$を式(3)で更新を行う.

$$UCB(s,a) = Q(s.a) + C\sqrt{\frac{ln\sum(N_a(s,:))}{N_a(s,a)}} \quad min\{\frac{1}{4}, V(s,a)+\sqrt{\frac{2ln\sum(N_a(s,:))}{N_a(s,a)}}\} \quad (3)$$

満たさない場合は0に更新を行う.また,Cは探索傾向を示す係数であり,本論文ではC=0.1で取り扱っている.その後$s=s'$に状態を更新しStep14に戻る.

一方,環境として(0.0,0.0)から(1.0,1.0)までの正方形の中と制約があり,それ以上の座標へ遷移することはできない.これを到達不可能状態と呼ぶ.また,従来手法では方向と出力の値に応じて決定される加速度ベクトルが速度ベクトルVに加えられ,座標Xが更新される.加速度をA,減衰率をζ(=0.1)としたときに式(4)式(5)となる.

$$V(t+1) = \zeta V + A(t) \quad (4)$$

$$X(t+1) = X(t) + V(t+1) \quad (5)$$

3 提案手法

従来手法における局所解に陥る原因として挙げられるのは,二つが大きく影響している.

一つ目はQ更新である.アルゴリズムにおける式(1)に該当する部分はゴール後に更新するQ学習とは違い逐次的に更新が行われ,部分報酬やゴール報酬を用いて状態更新の度に更新し最適化を図っていく.しかし,従来手法では部分報酬が存在せずゴール報酬のみである.これが原因となり,Q値更新の際に,状態がエージェントから遠ければ遠いほど報酬が極端に小さくなり,エージェントとゴールが離れた位置に置かれると学習不十分状態で行動をとることになる.

二つ目は行動選択方法にある.従来手法での選択はgreedy手法が採用されており,最も評価値の高いUCB値が行動を決める.この際にUCB値の変動するのはStep10に入る条件であるStep9の状態遷移が起こったことに起因する.つまり,状態遷移が起こらなければUCB値は変動しない.例えば次の行動を取る際に到達不可能状態が最も評価値の高いUCB値だった場合,到達不可能状態に行動をするが,何回繰り返したとしても更新は起きない.更新が起きなければ,UCB値は変動することがないために無限にループすることとなる.

一つ目は二つ目の現象が起こることを助長し,二つ目は局所解に陥るその原因である.これを取り除くためにそれぞれにおける方法を提案する.一つ目に対する改善案は部分報酬が無いことに起因する.そのため探索段階において,ゴール以外での報酬を設けることにした.この探索段階で報酬を与えることで,学習段階でのエージェントから遠い状態での学習改善を期待する.その報酬として,探索段階ではゴール後に式(1)を利用した式(6)を使い,図1のフローチャートの方法に従う.

$$PR(s,a) = PR(s.a) + \alpha[R_t + \gamma max_{a'\in A(s')}PR(s',a') - PR(s,a)] \quad (6)$$

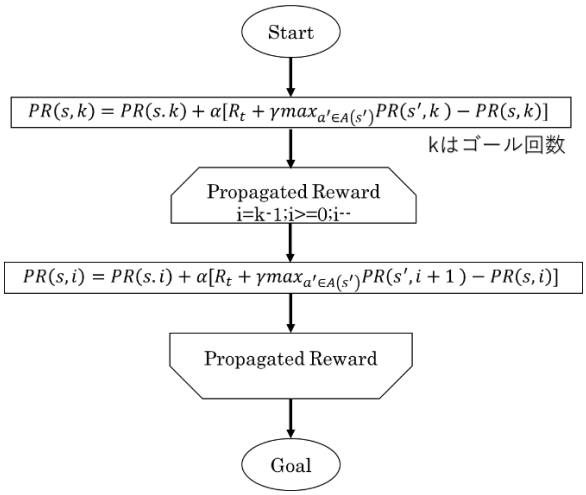

図1 伝搬Q学習のフローチャート

ゴールから遡りQ学習を行う方法と従来手法の逐次型のQ学習を区別するために,伝搬報酬を用いたPRQ学習とする.また,Step4におけるループを抜けた後に図1の処理を行う.ゴール後,次エピソード時の(1)式の更新の代わりに式(6)のように評価値を部分報酬として用いる更新式(7)にする.

$$Q'(s,a) = Q(s.a) + \alpha[R_t + PR(s,a) + \gamma max_{a'\in A(s')}Q(s',a') - Q(s,a)] \quad (7)$$

また,探索段階のみこの更新式は適用されるのでそれ以外は(1)式に順守する.

二つ目に対する改善案は式(3)が状態更新の際にしか変動しないことに依るため,状態更新時以外の要素を取り入れる必要がある.そこで,状態更新回数N_aだけでなく行動選択回数NN_aを取り入れる.以下がその式(8)である.

$$UCB(s,a) = Q(s.a) + C\sqrt{\frac{\ln\sum(N_a(s,:))}{N_a(s,a)}\min\{\frac{1}{4}, V(s,a) + \sqrt{\frac{2\ln\sum(NN_a(s,:))}{NN_a(s,a)}}\}\sqrt{\frac{\ln\sum(NN_a(s,:))}{NN_a(s,a)}}} \quad (8)$$

式(8)を利用することで一度到達不可能状態に遷移しようとしても繰り返すことでUCB値は低下していくことになる.そうすることで,少しずつであるが,自ら局所解から抜け出すことができると期待する.これを伝搬Q学習も利用することでより安定した結果の取得を期待してPRDUCBQ学習とする.

4 実験結果および検討

提案手法で挙げた二つの提案を挙げた.一つ目を方式Aとし,二つ目を方式Bとする.また,方式Aと方式Bの両方を用いたものを方式Cとする.また,これらは試行回数二十回の平均データを用いた.以下に方式Aと方式Bの比較した結果を表1,図2に表す.

表1　方式Aと方式BのEp900~1000

	方式A	方式B
平均ゴールターン	134.53	128.69
分散	1301.77	909.78

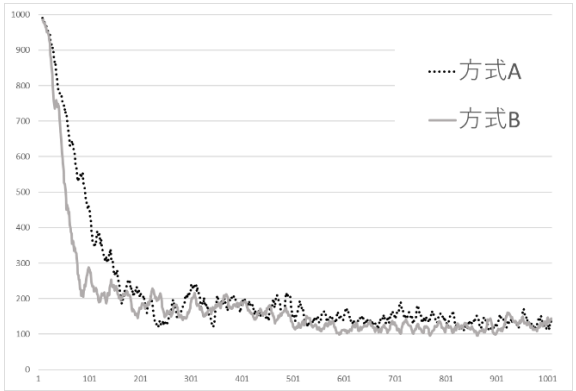

図2　方式Aと方式Bの平均ゴールターン

表1から方式Aと方式Bのエピソード数が900回から平均と分散を取った値を比較してみるとゴールまでの平均ゴールターンは方式Bが小さく,分散値も同様に小さく安定している.また図2からは探索段階においての学習速度も方式Bが上回っていることがわかる.ここから次に方式Bと方式Cの比較をする.以下に方式Bと方式Cの比較した結果を表2と図3に表す.

表2　方式Bと方式CのEp900~1000

	方式B	方式C
平均ゴールターン	128.69	101.37
分散	909.78	191.61

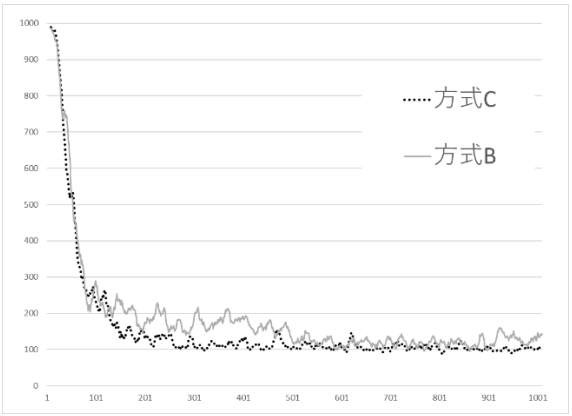

図3　方式Cと方式Bの平均ゴールターン

表2から表1と同様に比較をすると,平均ゴールターンは方式Bが20回程方式Cより小さく,分散は方式Bが大きく安定している.また図3から方式Bと方式Cは後半のエピソード回数ではともにゴールまでの行動数は100回に近い.ただ早期収束する方式は,エピソード数300回において収束した行動数に近い値を結果として出している方式Cが優良であると言える.これは方式Aの特徴である探索段階での部分報酬が大きく影響している.本論文は安定化を図っているため早期安定化させていることには大きな価値があるとして従来手法と比較するのは方式Bではなく方式Cを用いる.以下に従来手法と方式Cの比較した結果を表3と図4に表す.

表3　方式Cと従来手法のEp900~1000

	従来	方式C
平均ゴールターン	142.89	101.37
分散	1438.49	191.61

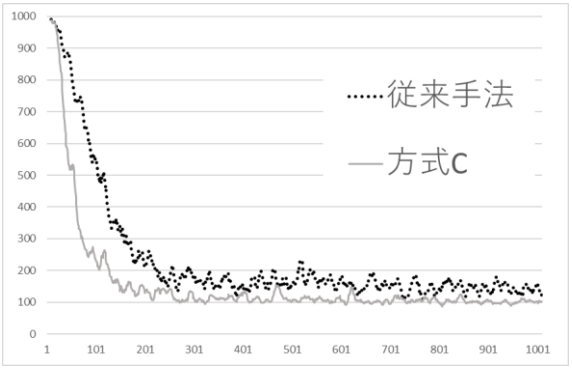

図4　従来手法と方式Cの平均ゴールターン

表3から表1と同様に比較すると,平均ゴールターンと分散ともに方式Cが小さくなっている.図4からも全体的によくなっていることがわかる.

5 まとめ

局所解を避けるための方法を提案し,具体的に複数の方式を挙げ平均値と分散値を上げることで比較した.結果,大きく変わったのは分散値であり,それに付随して平均値も改善された.方式Aは単体では改善されたとは言えなかったが,方式Bと合わせた方式Cとすることで方式Cは少ない学習回数で方式Bより早くデータを収束させることを可能とした.また,方式Cは大きく安定した分散値を獲得しており,本研究の課題点を解決し,平均ゴールターンが向上したと捉えることができる.しかし,エージェントの動きとして到達不可能状態側を沿う形でゴールへと向かう動きが見られる.ゴールすることは可能だが,敢えて遠回りをしている場面が多々あり,この現象が課題となる.この問題を解決することができれば平均ゴールターンはより向上することが期待できる.

実験環境として「強化学習におけるUCB行動選択手法の効果」が原案となっており2),IEEE Car Rasing Competitionと言われるように自動車がモデルとなっている.しかし,減衰率ζが0.1となっている本環境では自動車本来の挙動とは言いづらい.そのため,減衰率ζの値を大きくした際にも上手く学習するのかも検討したい.

【参考文献】

1) 野津　亮,本多　亮宏,Discounted UCB1-tuned の Q 学習への適用,知能と情報(日本知能情報ファジィ学会誌)Vol.26,No6, (2014)pp.913-923

2) 斎藤　晃貴,野津　亮,本多　亮宏,強化学習における UCB 行動選択手法の効果,30th Fuzzy System Symposium (Kochi, September 1-3, 2014) pp174-179