平成 27 年度　情報工学科卒業研究概要

成功確率と収益を組み合わせた価値関数に基づく強化学習

犬塚・武藤研究室　知能系
No. 24115133　堀江 直人

1　はじめに

強化学習の手法の 1 つに，竹山らが提案した，成功確率に基づく強化学習 [1] というものがある．成功確率に基づく強化学習は，各状態行動対の成功確率の期待値を求め，成功確率の低い行動を行動選択から取り除き，安全な行動選択を行う手法である．しかし，成功確率の低い行動を行動選択から取り除くため，リスクが回避不可能な場合の行動選択を獲得することは不可能である．

そこで，本論文では成功確率と従来の行動価値を組み合わせた新しい価値に基づく手法を提案し，リスクが回避可能な場合に加え，リスクが回避不可能な場合でも適切な行動を学習することを示す．

2　従来手法

成功確率に基づく強化学習は，従来の報酬の代わりに，行動が成功したか失敗したかを報酬とし，時刻 t において成功した場合 $r_{t+1} = 1$，失敗した場合 $r_{t+1} = 0$ を与える．

竹山ら [1] は状態 s において行動 a を選択したときの価値関数 $\Pr Q(s,a)$ を

$$\Pr Q(s,a) = E_\pi \left[\log \prod_{k=0}^{\infty} P(r_{t+k+1})^{\gamma^k} \right]$$

で定義する．$P(r_{t+k+1})$ は時刻 $t+k$ での成功確率を表す．E_π は方策 π の下での成功確率の期待値である．竹山らは状態 s で行動を選択するときに $\Pr Q(s,a)$ が一定の値より低い行動を選択肢から取り除き，残った行動の中から行動を選択することで，安全な行動を選択をする手法を提案した．

3　提案手法

従来手法は成功確率の低い行動を選択肢から取り除くことで安全な行動を選択させたが，すべての行動にリスクが存在する環境では，行動選択の方策を獲得することは不可能である．

そこで，本論文ではリスクを回避できない環境での行動選択を学習させるための手法として，成功確率と収益を組み合わせた価値関数を提案する．本研究での価値関数を EQ とする．

$$Q(s,a) = E_\pi \left[\sum_{k=0}^{\infty} \gamma^k (r_{t+k+1}) \right]$$

$$EQ(s,a) = \begin{cases} e^{\Pr Q(s,a)} \times Q(s,a) & (Q(s,a) \geq 0) \\ (1 - e^{\Pr Q(s,a)}) \times Q(s,a) & (Q(s,a) < 0) \end{cases}$$

$EQ(s,a)$ により成功確率が低い行動の価値を小さくしつつも，竹山らの手法とは異なり，選択肢から除外しないようにすることができる．

4　実験

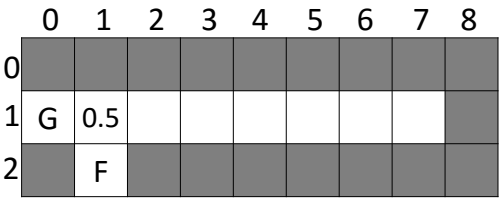

図 1: 格子世界

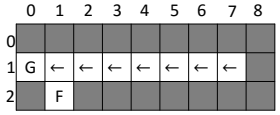

図 2: 提案手法

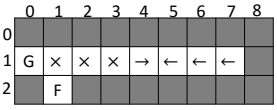

図 3: 竹山らの手法

リスクを回避できる経路がある格子世界と，図 1 に示すリスクを回避できない格子世界を用意し，エージェントの行動についてそれぞれ調査した．格子世界には，終端状態として，G，F が存在する．灰色のマスは移動不能なマスとして扱う．G に到達した場合，エージェントは報酬を得る．その後次のエピソードを実行する．F のマスに到達した場合，そのエピソードは失敗とし，報酬を得た後に，次のエピソードを実行する．$(1,1)$ から G に移動しようとした場合，0.5 の確率で，F のマスに移動する．

リスクの回避できる環境では，竹山らの手法と同様に提案手法はリスクを回避する経路を学習した．一方でリスクを回避できない環境では，提案手法は，G に向かう経路を学習した (図 2) が，竹山らの手法では，G に到達することができなかった (図 3)．

5　まとめ

本論文では，竹山らが提案した成功確率に基づく強化学習手法の価値関数 $\Pr Q(s,a)$ を基に，成功確率と収益を組み合わせた行動価値関数 EQ を提案した．これにより，リスクを回避する経路を学習することができ，また，リスクを回避することができない場合でも，目標状態に到達できる行動を学習させることができた．

参考文献

[1] 竹山大貴, 加納政芳, 松井藤五郎, 中村剛士: 成功確率に基づく強化学習によるロボットの危険回避行動の獲得, 知能と情報, vol.27, no.6, pp.877-884, 2015.