1M-03

POMDPs環境における強化学習を用いたロボットの衝突回避行動学習

田中健太　長名優子

東京工科大学 コンピュータサイエンス学部

1　はじめに

教師信号を用いずに環境との相互作用により適切な行動を行うための政策を獲得するための学習方法として，強化学習に関する様々な研究が行われている [1]．不完全知覚状態が存在する部分観測マルコフ決定過程 (POMDPs： Partially Observable Markov Decision Processes) 環境下で決定的政策を学習する方法として，POMDPs 環境のための決定的政策を学習する Profit Sharing[2] が提案されている．

本研究では，POMDPs 環境における強化学習を用いたロボットの衝突回避行動学習を実現する．

2　Profit Sharing

Profit Sharing[3] はエージェントの観測と行動の組をルールとし，報酬を基にルールの価値を更新することで学習を行う．エージェントが報酬を獲得したときに，初期状態から報酬を得るまでの一連のルール (エピソード) に報酬を以下のように分配する．

$$q(o_x, a_x) \leftarrow q(o_x, a_x) + r \cdot F(x) \tag{1}$$

ここで，$q(o_x, a_x)$ は時刻 x における観測 o_x のときに行動 a_x を取るというルールの価値，r は報酬量を表し，以前のルールの価値に強化関数 $F(x)$ に基づいて分配された報酬を加算することで価値を更新している．強化関数 $F(x)$ は

$$F(x) = \frac{1}{(|C^A| + 1)^{W-x}} \tag{2}$$

で与えられる．ここで，$|C^A|$ はエージェントの取りうる行動の数，W はエピソードの長さ，x は時刻を表す．報酬獲得の直前のルールに最も多く報酬が分配され，報酬獲得時の時刻から離れるほど分配される報酬の量が減るようになっている．

行動選択にはボルツマン選択を用いる．観測 o のときに行動 a を取る確率 $P(o, a)$ は

$$P(o, a) = \frac{\exp(q(o, a)/T)}{\sum_{b \in C^A} \exp(q(o, b)/T)} \tag{3}$$

で与えられる．ここで，T は温度パラメータであり，時間経過とともに 0 に近づけていく．また，C^A はエージェントが取りうる行動の集合である．T の値は学習の開始直後では大きな値に設定されるため，行動はほぼランダムに選択される．学習が進むにつれて T の値は 0 に近づくため，価値の高いルールの行動が高確率で選択されるようになる．

3　POMDPs環境のための決定的政策を学習する Profit Sharing

エージェントは環境から観測を得る．このとき，過去に観測が不完全知覚状態と判断されていれば過去の観測の履歴を用いて行動選択を行う．不完全知覚状態と判断されていない場合，現在の観測のみを用いて行動選択を行う．行動を行った結果，報酬が得られていなければ再び環境から観測を受ける．報酬が得られていれば，エージェントはまずそのエピソードで報酬獲得のために決定的な政策をとることができているかをチェックする．その後，決定的な行動選択を行えていないような観測を決定的な政策をとるために情報が足りていないと判断し，不完全知覚状態として判定する．さらに得られた報酬を元に学習を行い，環境をリセットし，最初から行動を行う．

3.1　行動選択

観測が不完全知覚状態と判断されていない場合，現在の観測に関するルールの価値に基づいて行動を決定する．時刻 x における観測 o_x での行動 a を選択する確率 $P(o_x, a, x)$ は

$$P(o_x, a, x) = \frac{\exp(q_n(o_x, a)/T(o_x))}{\sum_{b \in C^A} \exp(q_n(o_x, b)/T(o_x))} \tag{4}$$

で与えられる．ここで，$T(o_x)$ は温度パラメータである．また，C^A はエージェントが取りうる行動の集合

Obstacle Avoidance by Reinforcement Learning in POMDPs Environment
Kenta Tanaka and Yuko Osana (Tokyo University of Technology, osana@stf.teu.ac.jp)

である．$T(o_x)$ の値は学習の開始直後では大きな値に設定されるため，行動はほぼランダムに選択される．学習が進むにつれて $T(o_x)$ の値は 0 に近づくため，価値の高いルールの行動が高確率で選択されるようになる．また，観測が不完全知覚状態と判断されている場合は，過去の観測の履歴を考慮して行動を決定する．このとき，行動選択確率は

$$P(o_x, a, x) = \frac{\exp(q_n(o_x, a)/T(o_x)) + Q(o_x, a, x)}{\sum_{b \in C^A} \exp(q_n(o_x, b)/T(o_x)) + Q(o_x, a, x)} \tag{5}$$

で与えられる．ここで，$Q(o_x, a, x)$ は過去の観測の系列を考慮したルールの価値の和である．

3.2 学習

POMDPs 環境のための決定的政策を学習する Profit Sharing では，観測が不完全知覚状態と判断されていない場合，時刻 x における観測に関するルール (o_x, a_x) の価値のみを更新する．このとき，ルールの価値は

$$q(o_x, a_x) \leftarrow \left(1 - \frac{1}{I(o_x, a_x)}\right) q(o_x, a_x) + \frac{r \cdot F(o_x)}{I(o_x, a_x)} \tag{6}$$

のように更新される．また，観測が不完全知覚状態と判断されていない場合には，時刻 x における観測に関するルール (o_x, a_x) の価値に加え，時刻 x における行動決定に関わった過去の観測の系列 $(O \rightarrow o_x)$ に関するルールの価値を更新する．このとき，ルールの価値は

$$q(O \rightarrow o_x, a_x) \leftarrow \left(1 - \frac{1}{I(O \rightarrow o_x, a_x)}\right) q(O \rightarrow o_x, a_x) + \frac{r \cdot F(O \rightarrow o_x)}{I(O \rightarrow o_x, a_x)} \tag{7}$$

のように更新される．ここで，$I(o_x, a_x)$ はルール (o_x, a_x) の強化回数，r は報酬，$F(o_x)$ は強化関数である．強化回数が少ないほど，値が大きくなる．

4 POMDPs 環境における強化学習を用いたロボットの衝突回避行動学習

本研究では，POMDPs 環境のための決定的政策を学習する Profit Sharing を用いることで，不完全知覚状態の存在する POMDPs 環境におけるロボットの衝突回避行動学習を実現する．

本研究では，図 1 のような障害物の存在する環境を設定し，実験を行う．ロボットは，取り付けられたセンサやカメラにより障害物や壁を観測する．ロボットは観測に基づいて直進，左への方向転換，右への方向転換のいずれかの行動を選択する．学習の開始地点から目標地点までを 1 つのエピソードとし，開始地点から目標地点までのステップ数が少ないほど大きな報酬を獲得できる．障害物や壁に衝突したら負の報酬を獲得するように設定する．学習が進むにつれて，ロボットは障害物や壁に衝突することなく，適切な行動選択を行えるようになる．

図 1: 実験環境

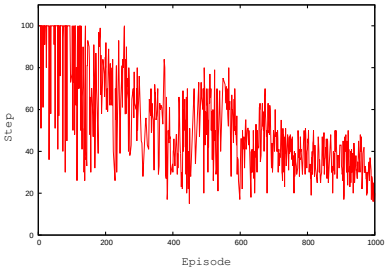

図 2: ステップ数の推移

5 実験

ロボットを用いた実験，およびシミュレーションによる実験を行った．図 2 にシミュレーションにおける試行ごとのステップ数の推移を示す．試行が進むにつれステップ数が少なくなっており，学習が行われていることが分かる．

参考文献

[1] R. S. Sutton and A. G. Barto : Reinforcement Learning : An Introduction, The MIT Press, 1998.

[2] Y. Takamori and Y. Osana : “Profit sharing that can learn deterministic policy for POMDPs environments,” Proceedings of IEEE International Conference on System, Man and Cybernetics, Anchorage, 2011.

[3] J. J. Grefenstette : “Credit assignment in rule discovery systems based on genetic algorithms,” Machine Learning, Vol.3, pp.225–245, 1988.