SSI 2017 に参加して

辻　　和輝
Kazuki TSUJI
電子情報学専攻修士課程　2 年

1. はじめに

2017 年 11 月 25 日から 27 日に静岡県の静岡大学浜松キャンパスで開催された計測自動制御学会　システム・情報部門　学術講演会 2017（SSI 2017）に参加し，「強化学習 Profit Sharing における行動選択に関する一考察」という題目でポスター発表を行った．

2. 研究背景

現在の工場では加工作業や搬送作業をロボットにより自動化させる Factory Automation 化（FA 化）が進んでいる．FA 化によって不良品の発生や，生産機械の操作ミスによるヒューマンエラーを低減することができ，作業工程の効率化が期待されている．また，市場や顧客のニーズの変化により大量生産から少量多品種の生産形態へ移行しており，工場内で作業を行うロボットには柔軟な対応が求められる．一方工場内での製品の加工やピッキング作業を行う多関節アームロボットはベルトコンベア上を流れる製品や，保管棚内の製品を掴む際に，自己位置や製品の位置を技術者があらかじめ教示しておき，固定的な軌道で作業を行うことが多い．少量多品種生産では生産する製品の変更や加工機械の入れ替えが行われることがあり，その際に熟練した技術者以外の作業者がアームロボットの軌道の教示を行った場合，熟練した技術者と比較して技術差が生じることが考えられる．

そのため，本研究では強化学習を教示に対して適用することで熟練者以外の作業者が教示を行った場合でも熟練者に近いレベルの教示を行えるようにすることを目的とする．

3. 強化学習

強化学習とはある環境内のエージェントが試行錯誤による報酬獲得を通じて行動選択の良し悪しを評価する方法である．エージェントはある時刻 t における状態 s_t を環境から観測し，状態 s_t から選択可能な行動 a_t を選び，実行する．行動を実行することで環境は次状態へと遷移し，エージェントは報酬値 r_t を受け取る．報酬値には次状態が目標状態であれば正の値を与え，目標状態でない場合は負，または 0 を与え，エージェントは報酬値を用いてルール（s_t，a_t）の価値 Q の更新を行う．

次状態へ遷移する際の行動を選択する際にいくつかの方法がある．ε-Greedy 選択では確率（1-ε）で価値が最大値である行動を選択し，確率 ε でランダム選択を行う．学習の初期は適度に環境内を探索する必要があるため確率 ε でのランダム選択を用い，学習が進むと一番価値が高い行動が最適行動となるため ε を 0 に近づけていく．ある環境において報酬獲得に 2 つ以上の行動選択が必要な場合，ε-Greedy 選択では適切に報酬を獲得できなくなる．その様な場合にはソフトマックス行動選択法であるルーレット選択を用いることが多い．ルーレット選択は獲得した報酬値の重みで行動を選択する方法であり，学習が進むにつれて報酬を獲得したルールの行動選択確率が高くなる．ルーレット選択ではあるルール（s,a）の行動選択確率 π（s,a）を次式で決める．

$$\pi(s,a_i)=\frac{Q(s,a_i)}{\sum_{j=1}^{N}Q(s,a_j)}$$

また，実ロボットで強化学習を行う場合には，試行回数が多い場合ロボットの耐久性や，作業時間の確保といった問題点があるため，収束速度の速い強化学習方法が必要とされる．

4. Profit Shairng

Profit Sharing ではルールの価値 Q の更新に報酬

獲得までの行動系列を遡り分配を行う．報酬の分配の際に，分配関数として等比減少関数 $f(x)$ を用いることで，報酬獲得に貢献しない無駄な経路の強化を抑制できることが知られている[1]．Profit Sharing の報酬割り当て式には次式が用いられる．

$$\omega(s_x, a_x) \leftarrow \omega(s_x, a_x) + r \times f(x)$$
$$Q(s_t, a_t) \leftarrow \omega(s_t, a_t)$$

5. 報酬割り当てによる学習性能の変化

Profit Sharing においてルーレット選択を用いた場合，割り当て報酬を Q 値に累積し続けていくため，学習が進むごとに各ルールの価値の差が大きくなり，報酬を獲得し続けるルールの行動選択確率が高くなっていく．

一方 Q 値を獲得推定報酬値に近づけていく更新方法を用いた場合は各ルールごとの差が出にくく，行動選択確率に影響が出ることが考えられる．ここで更新型の Profit Sharing として次式を用いる．

$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha\omega(s_t, a_t)$$
α：学習率

学習の方法として累積型と更新型の Profit Sharing を用い，行動選択方法としてルーレット選択を用いて迷路走行[2]（Figure 1）を行う．また，Q 値の初期値に対して報酬値の大小により，学習速度の問題[3]が起きるため，ここでは初期値を大中小の 3 パターン与える．結果を Figure 2 に示す．

6. まとめ

本研究では Profit Sharing における累積型報酬分配と更新型報酬分配の学習性能の変化を比較した．更新型では Q 値の初期値に依存せずに学習が進むことが分かった．これにより熟練者以外の教示結果に対してでも学習を行えることが期待できる．

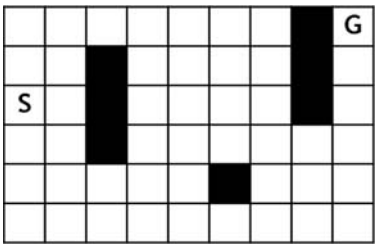

Figure 1　迷路

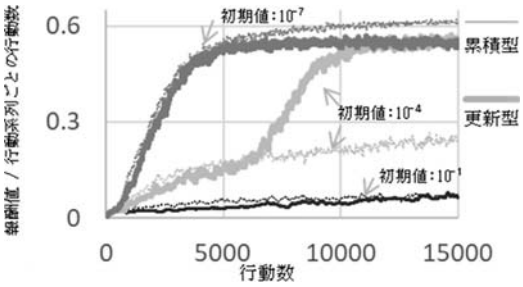

Figure 2　結果のグラフ

最後に研究発表を行うにあたり，ご指導いただいた植村先生，植村研究室の皆様に感謝致します．

謝辞

本研究は JSPS 科研費 15K16313 の助成を受けたものである．

参考文献

1）宮崎和光，山村雅幸，小林重信“強化学習における報酬割当ての理論的考察”人工知能誌，Vol.9，No.4，pp.580-587, 1994

2）Sutton, R. S. and Barto, A. G.：“Reinforcement learning －an introduction－” the MIT Press, 1998

3）植村　渉，上野敦志，辰巳昭治，“経験に固執しない Profit Sharing 法，”人工知能論文誌，Vol.21，No.1，pp.81-93, 2006