5U-02

# POMDPs環境を考慮した畳み込みニューラルネットワークを用いたProfit Sharing

野村和貴　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

近年、深層学習と強化学習を組み合わせた手法である深層強化学習が注目を集めており、Q Learning[1]を畳み込みニューラルネットワーク [2] を用いて実現したDeep Q-Network[3] や畳み込みニューラルネットワークを用いたProfit Sharing[4] などが提案されている。しかしながら、いずれの手法においても適切な政策が学習できないような課題が存在することもわかっている。これは、学習しようとしているゲームに不完全知覚状態が含まれているためであると考えられる。不完全知覚状態が含まれている環境は部分観測マルコフ決定過程 (Partially Observable Markov Decision Processes: POMDPs) 環境と呼ばれる。それに対し、POMDPs 環境において決定的な政策を学習することができる手法として、 POMDPs 環境のための決定的政策を学習する Profit Sharing[5] が提案されている。この手法では、観測が不完全知覚状態であると判断されている場合には、現在の観測だけではなく観測の履歴を考慮して行動の選択を行うことで、 POMDPs 環境下でも決定的な行動選択が行えるような学習が実現されている。

本研究では、POMDPs 環境のための畳み込みニューラルネットワークを用いたProfit Sharing を提案する。

## 2 畳み込みニューラルネットワークを用いたProfit Sharing

畳み込みニューラルネットワークを用いた Profit Sharing[5] は、Profit Sharing の行動価値の学習を畳み込みニューラルネットワークで行うものであり、3層の畳み込み層と2層の全結合層から構成される畳み込みニューラルネットワークを用いる。誤差関数 $E$ はProfit Sharing の価値の更新式に基づいた以下のような2乗誤差で与えられる。

$$E = \frac{1}{2}\left(rF(\tau) - q(o_\tau, a_\tau)\right)^2 \tag{1}$$

ここで、$r$ は報酬、$q(o_\tau, a_\tau)$ は観測 $o_\tau$ において行動 $a_\tau$ をとることの価値を表す。$F(\tau)$ は時刻 $\tau$ における報酬分配の割合を表しており、報酬分配関数 $F(\cdot)$ としては

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

のような等比減少関数が用いられる。ここで、$|C^A|$ はエージェントのとり得る行動の集合、$W$ はエピソードの長さを表す。

## 3 POMDPs 環境を考慮した畳み込みニューラルネットワークを用いた Profit Sharing

本研究では、POMDPs 環境を考慮した畳み込みニューラルネットワークを用いたProfit Sharing を提案する。提案手

法では、入力として用いる観測の長さの異なる複数の畳み込みニューラルネットワークを用いる手法である。観測ごとの行動の決定度と学習の進行度を用いて不完全知覚状態の判断を行い、不完全知覚状態であると判断された場合には、より多くのステップにおける観測を入力として扱えるようなネットワークに切り替え、その出力によって行動選択を行うようにすることで、適切な行動を選択できるようにする。

### 3.1 構造

提案手法では、入力として用いる観測の長さの異なる複数の畳み込みニューラルネットワークを用いる。それぞれの畳み込みニューラルネットワークは文献 [5] で用いられているのと同様に3つの畳み込み層と2つの全結合層とから構成されている。入力として用いる観測の長さとしては様々なものが考えられるが、実験では4と8の2種類のネットワークを用いている。な

---

Profit Sharing using Convolutional Neural Network considering POMDPs Environment
Kazuki Nomura and Yuko Osana (Tokyo University of Technology, osana@stf.teu.ac.jp)



Copyright ©2020 Information Processing Society of Japan.
All Rights Reserved.

お、基本となるネットワークにおいて考慮する観測の長さは 4 とする。これは、ゲームを題材として学習を行う場合には、文献 [5] をはじめとして多くの研究において 4 フレーム分のゲーム画面を入力として用いられているためである。

### 3.2 不完全知覚状態の判断

提案手法では、観測ごとの決定度と学習の進行度を用いて不完全知覚状態であるかの判断を行う。

観測 $o$ における行動の決定度 $d(o)$ は

$$d(o) = \frac{\sum_{a \in C^A} \left( \frac{\max(0, q(o,a))}{\sum_{b \in C^A} \max(0, q(o,b))} - \frac{1}{|C^A|} \right)^2}{N} \quad (3)$$

で与えられる。ここで、$|C^A|$ はエージェントのとり得る行動の集合、$q(o,a)$ は観測 $o$ における行動 $a$ の行動価値を表している。また、決定度 $d(o)$ は観測 $o$ においてとり得る行動の行動価値の比率を用いて算出しているため、行動価値に負の値が含まれていると正しく比率を求めることができない。そのため、行動価値が負の値の場合は、0 として扱うことにする。また、$N$ は $d(o)$ の最大値が 1 になるような正規化定数であり、以下のように与えられる。

$$N = \frac{|C^A| - 1}{|C^A|} \quad (4)$$

式 (3) において、行動選択が決定的に行われている場合には、特定の行動に対する行動価値の値が大きく、それ以外の行動に対する行動価値の値が小さくなっていると考えられる。そのため、行動の決定度 $d(o)$ は 1 に近い値をとることになる。それに対し、行動選択がランダムに行われている場合には、行動の決定度 $d(o)$ は 0 に近い値をとることになる

提案手法では、学習が進行しているにも関わらず行動が確率的に選択されているようであれば不完全知覚状態であると判断する。学習の進行度は現在のステップ数 $t$ がしきい値を超えている場合に学習が進行していると判断する。学習の進行度と行動の決定度 $d(o)$ に基づいて

$$t > \theta^l \quad (5)$$

$$(1 - d(o)) > \theta^{PA} \quad (6)$$

が成り立つとき、観測 $o$ は不完全知覚状態であると判断される。ここで、$\theta^l$ は学習の進行度の判断に用いるしきい値、$\theta^{PA}$ は不完全知覚状態の判断に用いるしきい値である。また、学習済みのネットワークを用いる場合には、行動の決定度 $d(o)$ のみを用いて不完全知覚状態の判断を行う。

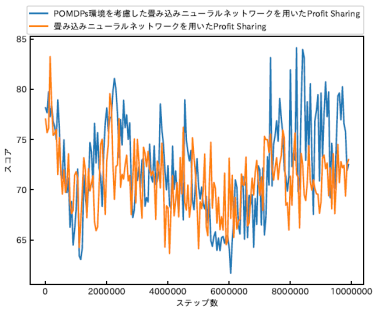


図 1: スコアの推移 (Asterix)

### 3.3 行動選択

不完全知覚状態と判断されていない観測では、4 フレーム分の観測を入力とする畳み込みニューラルネットワークの行動価値を用いて行動選択を行う。それに対し、不完全知覚状態と判断された観測では、 8 フレーム分の観測を入力とする畳み込みニューラルネットワークの行動価値を用いて行動選択を行う。行動選択には $\varepsilon$ − greedy 法を使用する。

## 4 計算機実験

提案手法と従来の畳み込みニューラルネットワークを用いた Profit Sharing[4] を用いて学習を行い、比較を行った際のスコアの推移の例を図 1 に示す。

## 参考文献

[1] C. J. C. H. Watkins and P. Dayan : “Technical Note: Q-Learning,” Machine Learning, Vol.8, pp.55-68, 1992.

[2] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner : “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, Vol.86, No.11, pp.2278-2324, 1998.

[3] V. Mnih *et al.* : “Human-level control through deep reinforcement learning,” Nature, No.518, pp.529-533, 2015.

[4] K. Hashiba and Y. Osana : “Study of learning ability in profit sharing using convolutional neural network,” Proceedings of IEEE International Conference on Artificial Intelligence and Soft Computing, Zakopnae, 2019.

[5] 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.



Copyright ©2020 Information Processing Society of Japan.
All Rights Reserved.