

自己組織化特徴マップに基づいた時系列パターンのための 確率的連想メモリによる強化学習の実現

新妻純 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

強化学習で扱う問題環境のうちエージェントの知覚能力に限界がある問題環境を部分観測マルコフ決定過程 (POMDPs: Partially Observable Markov Decision Processes) 環境といい, POMDPs 環境で起こる問題を不完全知覚問題という. POMDPs 環境のための決定的政策を学習する Profit Sharing[1] を KFM (Kohonen Feature Map) 連想メモリ [2] によって実現することで, 観測にノイズが含まれるような環境下でも行動選択を行える手法 [3] が提案されている. しかし, この手法では強化学習による学習とニューラルネットワークの学習を同時に行えないという問題がある.

本研究では, 自己組織化特徴マップに基づいた時系列パターンのための確率的連想メモリを提案し, それを用いて POMDPs 環境のための決定的政策を学習する Profit Sharing[1] を実現する. この手法では, 強化学習による学習とニューラルネットワークの学習を同時に行うことができる.

2 自己組織化特徴マップに基づいた時系列パターンのための確率的連想メモリ

自己組織化特徴マップに基づいた時系列パターンのための確率的連想メモリでは, 時系列パターンを逐次的に学習することができる. 学習において, パターンを一つのニューロンに対応させることで記憶させるが, 信頼度をパターンごとに設定することが可能となっており, 共通項に対応する複数の時系列パターンを信頼度に応じた比率で想起することができる. さらに, 未学習のパターンが入力された場合に, ニューロンの追加を行い, 学習を行うことができる.

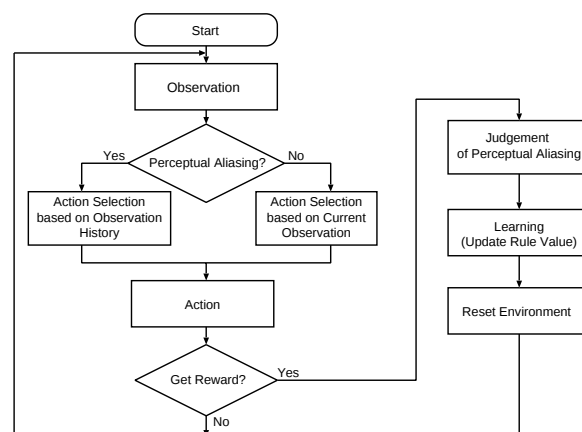


図 1: POMDPs 環境のための決定的政策を学習する Profit Sharing の流れ

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

POMDPs 環境のための決定的政策を学習する Profit Sharing では, POMDPs 環境において決定的な政策を獲得することを目的として学習を行う.

図 1 にこの手法の流れを示す. エージェントは環境から観測を受け取り, 観測が過去に不完全知覚状態であると判断されていれば過去の観測の履歴を考慮した行動の選択を行う. 不完全知覚状態であると判断されていない場合には, 受け取った観測のみを考慮して行動の選択を行う. 行動をとった結果, 報酬が得られなければ再びエージェントは環境から観測を受け取り, 行動の選択を行う. 報酬が得られていれば, 直前のエピソードで新たに不完全知覚状態として考慮しなければならない観測または観測の系列があるかどうかの判断を行う. 新たに不完全知覚状態として判定された観測または観測の系列があればそれらも以降のエピソードにおいて行動選択を行う際に利用されることになる. 次に, 直前のエピソードの情報を用いて観測または観測の系列に対するルールの価値の更新を行う. 環境は報酬が得られるたびにリセットされ, 次のエピソードが開始される.

4 自己組織化特徴マップに基づいた時系列パターンのための確率的連想メモリによる強化学習の実現

この手法では、POMDPs 環境のための決定的政策を学習する Profit Sharing[1] において観測や行動をパターンとして表現し、自己組織化特徴マップに基づいた時系列パターンのための確率的連想メモリを用いてルールを学習し、ルールの価値に応じた行動の選択を実現する。提案手法の流れは、基本的には図 1 と同じであるが、行動選択とルールの学習の部分で自己組織化特徴マップに基づいた時系列パターンのための確率的連想メモリを用いて実現している。

自己組織化特徴マップに基づいた時系列パターンのための確率的連想メモリは図 2 に示すように入出力層とマップ層から構成されている。自己組織化特徴マップに基づいた時系列パターンのための確率的連想メモリでは、入出力層の入力部にパターンが入力されるとマップ層を介して入出力層の出力部に次の時刻のパターンが出力される。そして、出力されたパターンを次の時刻に入出力層の入力部に入力するということを繰り返すことで履歴を考慮した時系列パターンの想起が実現されている。それに対し、POMDPs 環境のための決定的政策を学習する Profit Sharing では、不完全知覚状態であると判断されている観測に関しては、複数の観測の系列に対して 1 つの行動を選択する。そのため、POMDPs 環境のための決定的政策を学習する Profit Sharing において獲得されるルールを学習する自己組織化特徴マップに基づいた時系列パターンのための確率的連想メモリでは、入出力層は観測または観測の系列が入力される入力部とそれに対応する行動が出力される出力部、そして観測または観測の系列と行動の対で表されるルールの価値を信頼度として表す部分から構成されている。入出力層の入力部に行動選択で考慮される観測または観測の系列すべてが入力として与えられ、その時刻にとる行動が出力部に出力されることになる。

5 計算機実験

最短で 22 ステップでゴールに到達することのできるような POMDPs 環境において、提案手法と従来の POMDPs 環境のための決定的政策を学習する Profit Sharing[1] を用いて学習を行ったときのステップ数の遷移を図 3 に示す。これは 100 回試行の平均であり、横軸はエピソード数、縦軸は報酬が得られるまでのステップ数を表している。図 3 より、提案手法におい

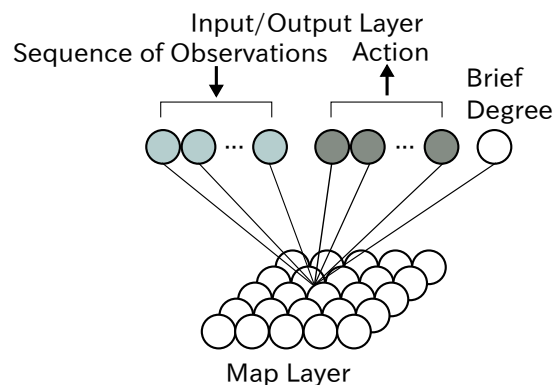


図 2: POMDPs 環境のための決定的政策を学習する Profit Sharing によって獲得されるルールの学習を行う自己組織化特徴マップに基づいた時系列パターンのための確率的連想メモリの構造

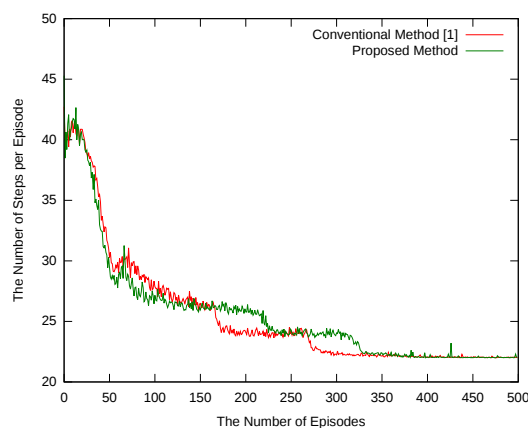


図 3: ステップ数の遷移

ても、POMDPs 環境のための決定的政策を学習する Profit Sharing と同様に最終的に最小ステップ数である 22 ステップで報酬が得られるように学習が収束していることがわかる。なお、提案手法ではノイズを付加したパターンを入力として与えて学習を行っている。

参考文献

- [1] 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.
- [2] T. Sato and Y. Osana : “Variable-sized Kohonen feature map probabilistic associative memory for sequential patterns,” Proceedings of IASTED Artificial Intelligence and Applications, Innsbruck, 2013.
- [3] D. Koma and Y. Osana : “Profit Sharing that can Learn Deterministic Policy for POMDPs Environments by Kohonen Feature Map Associative Memory,” Proceedings of IEEE International Conference on System, Man and Cybernetics, Manchester, 2013.