動的当たり確率の多腕バンディット問題における レーザカオスを用いた意思決定方式の性能向上

Improvement of decision making using laser chaos for multi-arm bandit problem with variable hit probabilities

埼玉大[1]，東京大[2]
○小田 章裕[1]，巳鼻 孝朋[1]，岩見 龍吾[1]，菅野 円隆[1]，成瀬 誠[2]，内田 淳史[1]
Saitama Univ.[1], The University of Tokyo[2]
○Akihiro Oda[1], Takatomo Mihana[1], Ryugo Iwami[1], Kazutaka Kanno[1], Makoto Naruse[2], and Atsushi Uchida[1]
E-mails: a.oda.652@ms.saitama-u.ac.jp, auchida@mail.saitama-u.ac.jp

はじめに: 近年、人工知能における強化学習の研究が盛んに行われている。その 1 つの問題としてバンディット問題[1]が挙げられる。これは当たり確率が未知である複数台のスロットマシンにおいて、複数回選択することにより報酬の最大化を図る問題である。

このバンディット問題を解く手法として綱引き理論が挙げられる[2]。綱引き理論とは、粘菌の摂食活動を模倣したアルゴリズムであり、報酬に応じてスロットマシンを選択する頻度を変える手法である。この綱引き理論を用いて光を用いた意思決定が多く提案されている。レーザカオスを用いた意思決定では、2 台のスロットマシンを仮定した時、可動なしきい値を設け、そのしきい値との比較により選択するスロットマシンを決定する。さらに、その報酬に応じてしきい値を変化させる方式である。この手法においてスロットマシンの台数の拡張が提案されている[3]。しかし、スロットマシンの当たり確率が動的に変化する問題における報告例は少ない。

そこで本研究では、当たり確率が動的に変化する多腕バンディット問題をレーザカオスを用いて解き、その性能向上を目的とする。

方法: 本研究では、戻り光を有する半導体レーザの時間波形を用いて、数値計算で意思決定を行う。2 台のスロットマシンを仮定し、可動なしきい値を設けた時、レーザカオスの時間波形としきい値の比較により選択するスロットマシンを決定する。さらにその報酬に応じてしきい値を変化させることでスロットマシンの選択頻度を変える。

スロットマシン i において当たった回数 H_i から試行回数 C_i を割ることで算出される推定当たり確率 $\widehat{P}_i$ を用いて、しきい値の移動量が決定される。本研究では記憶係数 β を導入し、以下の式で求める。

$$\mathrm{H_i}(t) = \begin{cases} 1 + \beta H_i(t-1) & \text{for win} \\ \beta H_i(t-1) & \text{for lose} \end{cases} \quad (1)$$

$$\mathrm{C_i}(t) = 1 + \beta C_i(t-1) \quad (2)$$

これにより過去の推定当たり確率の情報を捨てることで、動的に変化する当たり確率に対応することができる。

さらに、全てのスロットマシンの当たり確率を正しく推定するために、しきい値分解能 b の時間変化を導入する。b はしきい値の 1 回における移動量を示す。本研究では、はじめに b を大きい値に設定して 1 回におけるしきい値の移動量を小さくすることで、両方のスロットマシンを選択できるようにする。その後、ある試行回数にて b を小さくすることにより、報酬に対するしきい値の移動量を大きくする。

ここで本研究における評価方法として、平均正答率CDR(t)を用いる。これは当たり確率の高いスロットマシンを選択した割合の平均を示す。

結果: 本研究では記憶係数 β の値としきい値分解能 b の時間変化の導入による意思決定の比較をそれぞれ行った。

はじめに、記憶係数 β を導入した時の平均正答率を Fig. 1 の青線に示す。当たり確率が入れ替わった後において、β を導入した方式の方が従来法（緑線）よりも高い平均正答率を示している。これは、β を導入することにより、少しずつ過去の情報を捨てることで、当たり確率の変化に対応できるためである。

次に、しきい値分解能 b の時間変化を導入した時の平均正答率を Fig. 1 の赤線に示す。しきい値分解能 b を導入することにより、より正しい当たり確率の推定が行われるため、さらに高い平均正答率を示すことが分かった。

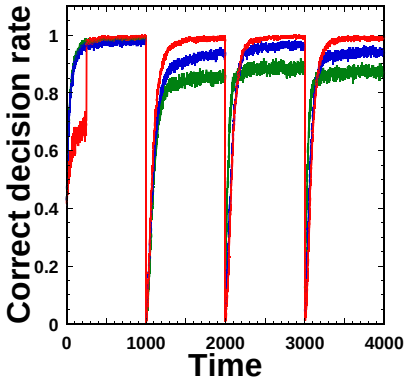

Fig. 1 Evolution of correct decision rate (CDR) for conventional method (green), the method with memory parameter $\beta = 0.986$ (blue), and the method with variable threshold resolution b (red)

まとめ: 本研究では、当たり確率が動的に変化する多腕バンディット問題において、記憶係数 β としきい値分解能 b の時間変化の導入による意思決定の性能向上を行った。本手法により、当たり確率の変化に応じた意思決定が実現できることが明らかになった。

参考文献

[1] 本多ら バンディット問題の理論とアルゴリズム 講談社 (2016).
[2] S. -J. Kim., et al., New J. Phys., **17**, 083023 (2015).
[3] M. Naruse., et al, Sci. Rep., **8**, 10890 (2018).