2Q-04

強化学習に基づくドローン配送問題のための報酬設計

村松俊輔† 林 冬惠‡
岡山大学工学部情報系学科† 岡山大学学術研究院環境生命自然科学学域‡

1.はじめに

近年，無人ドローンによる配送が注目されている．先行研究 [1]では，複数のドローンが限られた経路での配送経路を最適化する問題をドローン配送問題と定義し，マルチエージェント強化学習による手法が提案されている．

しかし，現在の報酬設計では取り扱う問題によって学習効果を保証できないため，適切な報酬設計が必要である．

本研究では，ドローン配送問題のマルチエージェント強化学習における学習効果を向上するため，各エージェントの行動とエージェント間の協調に関する報酬設計を行い，実験によって提案手法の有用性を示す．

2. ドローン配送問題

先行研究で提案されたドローン配送問題では，ドローンが移動する領域を図 1 のような無向グラフによって表現する [1]．各ドローンは互いに同じノードに存在することや，同じエッジ上を逆行及び同方向に進行することはできない．ドローンは環境から与えられたゴールを目指し，時間内の到着を目標に行動する．また，ドローン配送問題のマルチエージェント強化学習環境として MARL4DRP [2]が考案されており，ここで定義されたエージェントの報酬を表 1 に示す．この報酬設計では問題の規模によって学習時にエージェントのゴール到着率が低くなってしまうという問題がある．これはエージェント毎の報酬が疎になることと，エージェントの協調に対する報酬設計がないことに起因し，これを改善する設計が必要である．

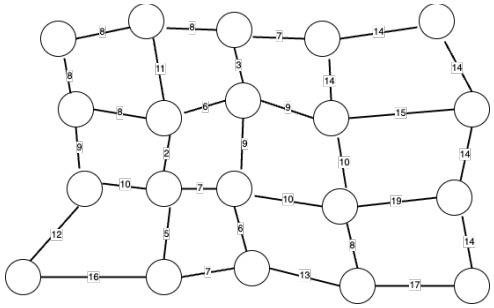

図 1 ドローン配送問題のイメージ図

表 1: ドローン配送問題の報酬

条件	報酬値
ゴール到着	100
他ドローンとの衝突時	-10
現在のノードで待機時	-10
別のノードへ遷移時	-1

3. 提案手法

本研究では 2 つの観点から報酬設計を提案する．

(1) ゴールに近づく行動に対する報酬設計

従来の報酬設計では，マップが大きくなるほどゴールまでの移動の負の報酬が大きくなってしまう．これを改善するため，ゴールに近づく行動に対して移動の負の報酬を半減し，この行動に動機を与える．エージェントごとに報酬が与えられ，ゴール率の底上げが期待される．

(2) 協調を促進する報酬設計

表 1 に示された報酬設計は，単に各エージェントに対する報酬に過ぎず，エージェント全体の協調に関する報酬設計が欠けている．そこで，本研究では flatland challenge [3]で提案されていた全体報酬を参考に，エージェントの協調報酬をドローン配送問題に適用する．具体的には，全てのドローンがゴールに到着時，これをドローンの協調の成果と見なし，ゴール報酬を 2 倍に増加させる設計を導入する．この設計により，全てのドローンが協調してゴールに到着することが促進される．

Reward Design for Reinforcement Learning in Drone Routing Problems

Shunsuke MURAMATSU†, Donghui LIN‡
†Department of Information Technology, Faculty of Engineering, Okayama University
‡Faculty of Environmental, Life, Natural Science and Technology, Okayama University

4.1. 評価

提案手法について，エージェント数は 3 として，ノード数 80 の 10x8 のマップを用いて実験を行った．マルチエージェント強化学習の手法は IQL [4]を使用する．ゴール率やタイムアップ率の指標を用いて評価を行った．

ゴール率とは全てのエージェントが衝突せず制限時間内にゴールできた割合のことである．タイムアップ率とは，エージェントがゴールせず，時間切れになった割合である．

4.2. 結果と考察

図 2~図 5 は提案手法の評価結果を示す．各図の横軸はその学習ステップ内のゴール率及びタイムアップ率の値である．評価結果が示すように 2 つの提案手法とも従来の手法に比べタイムアップ率が減少し，ゴール率の底上げに貢献している．提案手法(1)では，ゴールに近づく行動への移動ペナルティが軽減されることによって，エージェントはよりゴールへの到着に対して積極的に行動するようになったことに起因する．

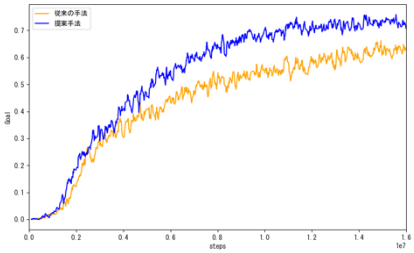

図 2 提案手法(1)の評価結果: ゴール率

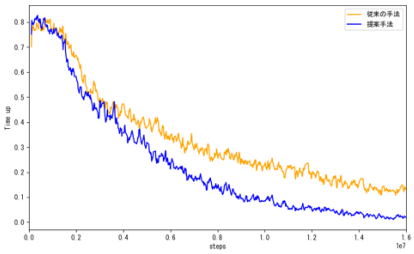

図 3 提案手法(1)の評価結果: タイムアップ率

提案手法(2)に関する効果はゴール率だけでなく，ゴール率の収束速度に対しても貢献をしており，学習ステップの少ない状態で高いゴール率を示している．これは，提案手法(2)による報酬が提案手法(1)と比べ，エージェントが得られる報酬値が大きく，ゴールに向かう行動をするエージェントがさらに増えたことに起因する．

この 2 つの提案手法の別のマップでの実験や，これらを組み合わせた手法に関する実験は今後行う予定である．

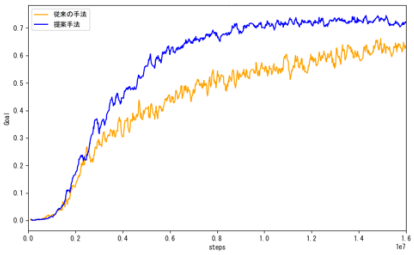

図 4 提案手法(2)の評価結果: ゴール率

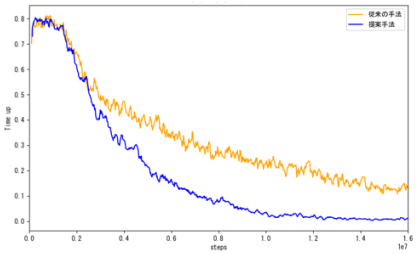

図 5 提案手法(2)の評価結果: タイムアップ率

5. おわりに

本研究では，ドローン配送問題のマルチエージェント強化学習においてエージェントのゴール到着への動機づけとエージェント間の協調促進に関する報酬設計を提案した．実験によって，既存手法と比べ，タイムアップ率が減少し，ゴール率が底上げされた．

今後は実世界をモデルにしたマップに対しても実験を行い，マップやエージェント数などの問題の条件ごとの適切な報酬設計について調査を行いたい．

謝辞

本研究は，日本学術振興会科学研究費基盤研究（B）（21H03556，2021 年度～2023 年度）の補助を受けた．

参考文献

[1] 青山秀紀, 丁世堯, 林冬惠. ドローン配送計画最適化問題のための最短経路情報を利用 したマルチエージェント強化学習. 第 36 回人工知能学会全国大会論文集, 2022.

[2] Shiyao Ding, Hideki Aoyama, and Donghui Lin. MARL4DRP: Benchmarking cooperative multi-agent reinforcement learning algorithms for drone routing problems. 20th pacific rim international conference on artificial intelligence, pp.459-465, 2023.

[3] Sharada Mohanty, et al. Flatland-rl: Multi-agent reinforcement learning on trains. arXiv preprint arXiv:2012.05893, 2020.

[4] Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In Proceedings of the tenth international conference on machine learning, pp. 330–337, 1993.