マルチエージェントシステムを用いた荷物搬送問題における強化学習の行動特性の検証

Verification of action selection of reinforcement learning
in the luggage transport problem using a multi-agent system

1265047　橋本　大輔　（Soft Intelligent System on a Chip 研究室）
（指導教員　星野　孝総　准教授）

1. はじめに

マルチエージェントシステムを用いた荷物搬送問題では，エージェントの数や環境に依存して学習状況が変化する．特に，エージェントの数が多いほど荷物搬送効率は向上するが，相互進路妨害が発生することが課題となる．また，エージェントの視界範囲が広い場合，学習の組み合わせが爆発的に増加するため，エージェントは部分観測で学習することが望ましい．しかし，部分観測による学習では，学習が正しく収束しないことが課題となっている．

本研究では，荷物搬送問題に強化学習[1]を用いることによるエージェントの行動特性の検証を目的とする．エージェントの行動特性を検証することで，環境に対する振る舞いの解析を目指す．また，部分観測環境下での正確な学習に向けて，相対ベクトルルールを導入した学習を提案する．

2. 相対ベクトルルールを導入した学習

強化学習によるエージェントの学習は，環境をもとに重み付き候補行動集合を生成するため，ルールはエージェントの視界環境に依存する．このルールを環境ルールと記す．しかし，部分観測におけるエージェントの学習は，環境同定の難しさから，環境ルールでは学習が正しく収束しない場合がある．また，ルールは荷物の有無で分けられる．

このため，本研究では環境に加え，現在位置とその 2 ステップ前との相対ベクトルをルールベースに導入する相対ベクトルルールを提案する．相対ベクトルルールを導入することにより，環境ルールでは学習が困難であった環境に対して，より多くのルールを生成できる．

環境ルールの$if-then$条件式を式(1)，相対ベクトルルールの$if-then$条件式を式(2)に示す．

$$if \ \text{環境} \ then \ \text{候補行動集合} \tag{1}$$

$$if \ \text{環境 and 相対ベクトル} \ then \ \text{候補行動集合} \tag{2}$$

相対ベクトルルールでは，相対ベクトルを条件に付加することにより，視界環境に依存する環境ルールより正確な学習が行えると考えられる．

3. 荷物搬送問題の評価実験

本研究において，強化学習におけるQ値と呼ばれるルールの有効価値を示す値の更新式として，堀内らによって提案された Q-PSPLearning[2]をもとにした更新式を用いた．

本研究では，強化学習を用いた荷物搬送問題の検証として，図 1 に示す実験環境でエージェントの学習を行った．エージェント数は 4 体であり，環境内で各エージェントは，上下左右と停止の 5 種類のいずれか行動を選択することで状態が遷移する．なお，外周は壁であり，各エージェントは他のエージェントと壁に衝突する行動選択は行えない．エージェントは，荷物搬入口から荷物を受け取り，荷物搬出口へ荷物を搬送することで報酬を受け取り，学習を行う．このため，荷物排出量が報酬の総和と等しくなる．また，各エージェントは独立しており，ルールを共有せず個別に学習を行う． 各エージェントの視界範囲は，自身を中心とした周辺 1 マスであり，壁と他のエージェントを区別できるが，荷物搬入口と荷物搬出口を知覚することはできない．

本実験では，学習を 500,000 ステップ試行し，学習終了までのエージェントの行動から評価を行う．

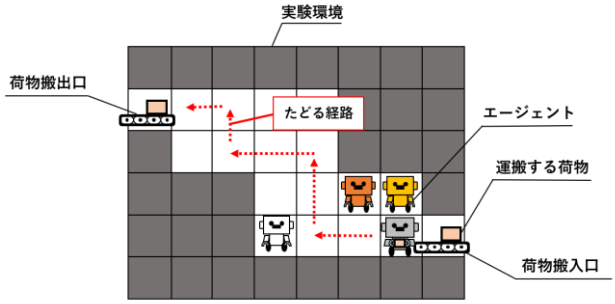

図 1. 荷物搬送問題の概要図

4. 結果と考察

環境ルールを用いた結果と相対ベクトルルールを導入した結果を図 2 に示す．環境ルールでは荷物搬送が安定しないが，相対ベクトルルールでは 1000 ステップあたり 120 回程度の排出量で安定している．相対ベクトルルールでは，3 体のエージェントが安定的に荷物搬送を行い， 残りの 1 体が停止・往来行動を続ける．

図 1 の実験環境では，両ルールともエージェントが 1 体が他のエージェントに挟まれるような状態となる．挟まれた場合，荷物搬送が行えないため，学習とQ値の更新が行えなくなる．この結果，環境ルールでは行動選択が安定せず，他のエージェントを阻害するため，荷物搬送が安定しなかった．

一方で，相対ベクトルルールを用いた場合，低いQ値でも相対ベクトルによる自己位置推定からその場にとどまり続け，阻害行動を選択しないと考えられる．この行動は，自身を犠牲とし，全体の利益を最大化させる利他的行動のように推察されるが，各エージェントは情報共有や他のエージェントの行動評価を導入していないため，強化学習の貪欲行動の結果であると考えられる．

この結果から，図 1 のような実験環境では，行動特性として他のエージェントを阻害する行動，利他的行動のように振る舞う行動を確認できた．

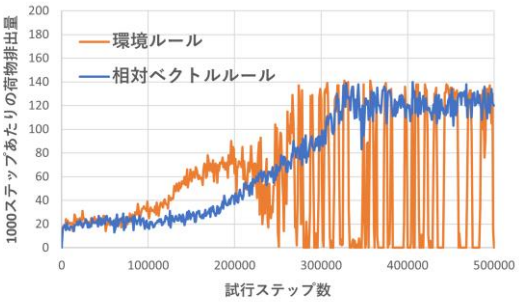

図 2. 環境ルールと相対ベクトルルールの実験結果

参考文献

[1] Sutton Richard S, Barto Andrew G: ”Reinforcement learning: An introduction,” MIT press, 2018.
[2]堀内匡,藤野昭典,片井修,椹木哲夫.”経験強化を考慮した Q-learning の提案とその応用,”計測自動制御学会論文集,Vol. 35,No. 5,pp 645-653,1999.