荷物搬送問題におけるデッドロック回避に向けた強化学習アルゴリズムのパラメータ設計と検証

1220130 橋本 大輔 （Soft Intelligent System on Chip 研究室）
（指導教員 星野 孝総 准教授）

1. はじめに

荷物搬送問題には、主に強化学習が問題解決に用いられている[1]。しかし、荷物搬送問題を強化学習で行う場合、デッドロックというエージェント同士の相互進路妨害が発生することがある。このデッドロックの発生が学習効率を著しく低下させるため、デッドロックの発生を抑制または回避が可能なエージェントの内部状態が求められる。

本研究では、荷物搬送問題におけるパラメータ設計によるデッドロック回避を目的とし、荷物搬送問題のもと実験を行った。

2. 実験手法

本研究では、堀内らによって提案された Q-PSPLearning[2]をもとにした式(1)を用いた。Q-PSPLearning は、強化学習の主な学習法である環境同定型学習に経験強化型学習の概念を導入したものであり、式(1)でQ値が更新される。Q値とはそのルールの有効価値を表す値である。ある時刻ステップtにおいて状態sにおける行動aを選択したときの価値関数は$Q(s,a)$となる。

$$Q(s_t,a_t) \leftarrow (1-\alpha)Q(s_t,a_t) + \alpha\left[r_t + \gamma \max_{a\in A} Q(s_{t+1},a_{t+1})\right] \quad (1)$$

$$r_t = rd^{t-n} \quad (2)$$

ここで、式(2)は報酬割り当てであり、式(2)におけるr_tは各ステップにおける報酬の値である。nは報酬獲得に費やしたステップ数である。また、dは報酬関数の公比であり、$0 < d < 1$である。

行動決定の方策は、式(3)で表されるボルツマン分布による確率的な行動選択を用いた。

$$\pi(s,a) = \frac{\exp\left(\frac{Q(s,a)}{T}\right)}{\sum_{b\in A}\exp\left(\frac{Q(s,b)}{T}\right)} \quad (3)$$

なお、Tは温度定数であり、この値が大きいほど各行動に対するQ値の差が行動選択に反映されなくなる。

強化学習のエージェントは一般に環境はマルコフ決定過程で行われる。しかし、本研究におけるエージェントは、環境全体を知覚できない不完全知覚で実験を行った。このため、部分観測マルコフ決定過程(POMDP)である。

3. 実験内容

本研究では、図 1 に示すような倉庫を模した 6×6 マスで荷物搬送問題を行った。マスの外周は壁であり、エージェントは他のエージェントと壁に衝突せず、壁の外へは出られない。エージェントは、自身を中心とした周囲 9 マスを観測でき、上下左右と停止の 5 種類の行動を選択できる。4 体のエージェントを荷物搬入口で荷物を受け取らせ、荷物搬出口まで運ばせるタスクを、すべてのエージェントが行動を選択し、状態が遷移するのを 1 試行とし 300,000 回試行を行った。この 300,000 回試行を 100 回行った場合の、式(2)における公比dを変化に伴う結果の変遷を確認した。公比dは 0.20、0.45、0.90 の場合に加えて、0.90 から 0.20 に減少させる手法と、0.45 時に荷物搬送が停止した場合に罰を与える手法を行った。

また、図 1 に示すエージェントの位置状態がデッドロックの発生例である。デッドロックが発生した場合、エージェント同士が相互進路妨害し、荷物排出量が著しく低下する。

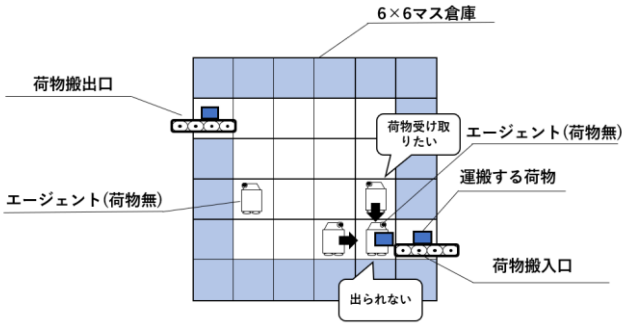

図 1 荷物搬送問題とデッドロックの概要図

4. 実験結果

実験を 300,000 回試行した場合、エージェントの行動は表 1 に示すように、公比にかかわらず 4 種類のパターンがみられた。最適収束 P1 は各エージェントが最適な行動を規則的にとり、収束 P2 は最適ではないが行動を規則的に行ったものである。収束しない P3 は規則的な行動を選択せず、荷物排出量一定とならないものであり、デッドロック P4 は荷物排出量が 0 となることが多いパターンである。

100 回実験を行った場合のパターンの数を表 1 に示す。表 1 から公比dは大きいほど収束しない P3 とデッドロック P4 が減少する。また、最適収束 P1 が増加する。罰を与えた場合、デッドロック P4 は減少するが、収束しない P3 が同値の公比と比べて増加した。

表 1 各公比で 100 回実験を行った結果

公比 d	0.20	0.45	0.90	0.90→0.20	0.45(罰)
最適収束 P1(回)	39	50	67	53	43
収束 P2(回)	33	29	27	28	31
収束しない P3(回)	20	9	2	6	25
デッドロック P4(回)	8	12	4	13	1

5. おわりに

本研究ではパラメータ設計によるデッドロック回避を目的として、公比の比較と罰を与える手法を試みた。その結果、公比の値は大きいほどデッドロックが発生しにくく、罰を与えた場合もデッドロックが抑制されることが確認できた。しかし、罰を与える手法の場合は収束しない結果が増加したため、適切な収束を行える罰の与え方を検討する必要がある。

参考文献

[1] 白川英隆,木村元,小林重信．”強化学習による協調的行動の創発に関する実験的考察,”知能シンポジウム資料,Vol. 25,pp.119-124,1998
[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.