エージェント工学 最終レポート課題

澤 祐里（21–1–037–0801）

2021 年 7 月 14 日

目次

1　目的

リサイクルロボットについての強化学習サンプルプログラムを用いて、パラメータを変更することなどで強化学習について考察する。

1.1　課題 1

リサイクルロボットの MDP を 3 種類作成する。

1.2　課題 2

ε Greedy 手法と Softmax 手法の概要を説明する。

1.3　課題 3

ε Greedy 手法と softmax 手法の比較を行う。

2 手順

2.1 課題 1

recycleRobotMDP.csv の表 1 を参考にして遷移確率や報酬を変更した case2, case3 を作成する。作成した報酬の与え方がエージェントの目標になっていることを示す。

表 1 元の MDP(case1)

状態	行動	次状態	遷移確率	報酬
START	N	high	1	0
high	search	high	0.8	12
high	search	low	0.2	4
low	search	GOAL	0.1	-3
low	search	low	0.9	2
high	wait	high	1	1
high	wait	low	0	1
low	wait	GOAL	0	1
low	wait	low	1	1
low	recharge	GOAL	1	0
low	recharge	low	0	0
GOAL	N	high	1	0

1. case2,case3 をそれぞれ以下のように作成した。

case2 表 2 のように、遷移確率は変更せず報酬だけを変更した。報酬の決め方は、以下の点を意識した。

- 行動「recharge」以外で状態「GOAL」になるということは、他の存在に助けてもらう必要がある。そのため、リサイクルロボット単体で完結するように、行動「recharge」以外で次状態が「GOAL」になる全ての行動の報酬を「-7」にした。
- 上記の理由で、行動「recharge」の報酬を「7」にした。
- リサイクルロボットに積極的に「search」してもらいたいため、行動「wait」の報酬を「-1」にした。

case3 表 3 のように、遷移確率を変更して、時間経過でリサイクルロボットのバッテリーが減るようにした。報酬の決め方は、以下の点を意識した。

- リサイクルロボットがバッテリーを無駄なく使えるように、バッテリーが減る場合の報酬を行動ごとにそれぞれ減らした。
- 行動「recharge」以外で状態「GOAL」になるということは、他の存在に助けてもらう必要がある。そのため、リサイクルロボット単体で完結するように、行動「recharge」以外で次状態が「GOAL」になる全ての行動の報酬を「-7」にした。
- 上記の理由で、行動「recharge」の報酬を「7」にした。
- リサイクルロボットに積極的に「search」してもらいたいため、行動「wait」の報酬を「-1」にした。

2. 以下に、引数「softmax 0.1 0.9 1000 recycleRobotMDP.csv S-RRlog.csv S-RRPolicy.csv S-RRQValue.csv」での、case2 と case3 の結果を示す。

case2 表 4 のような結果となった。

- 状態「high」の時、行動「search」の Q 値が行動「wait」よりも 2 倍以上高い。これは、case2 の報酬の与え方が、エージェントに積極的に行動「search」を行なわせていることを示している。

表 2　作成した MDP(case2)

状態	行動	次状態	遷移確率	報酬
START	N	high	1	0
high	search	high	0.8	10
high	search	low	0.2	4
low	search	GOAL	0.1	-7
low	search	low	0.9	2
high	wait	high	1	-1
high	wait	low	0	0
low	wait	GOAL	0	0
low	wait	low	1	-1
low	recharge	GOAL	1	7
low	recharge	low	0	0
GOAL	N	high	1	0

表 3　作成した MDP(case3)

状態	行動	次状態	遷移確率	報酬
START	N	high	1	0
high	search	high	0.8	10
high	search	low	0.2	4
low	search	GOAL	0.1	-7
low	search	low	0.9	2
high	wait	high	0.9	-1
high	wait	low	0.1	-4
low	wait	GOAL	0.1	-7
low	wait	low	0.9	-1
low	recharge	GOAL	0.8	7
low	recharge	low	0.2	-1
GOAL	N	high	1	0

表 4　case2 の結果

状態	行動	Q 値	Q 値更新数
GOAL	N	0	0
START	N	0	0
high	search	11.194368	54
high	wait	4.146994	54
low	recharge	4.310171	94
low	search	1.989767	94
low	wait	1.20334	94

表 5　case3 の結果

状態	行動	Q 値	Q 値更新数
GOAL	N	0	0
START	N	0	0
high	search	9.995437	58
high	wait	3.695292	58
low	recharge	3.895574	109
low	search	1.728605	109
low	wait	0.750725	109

– 状態「low」の時、行動「recharge」の q 値が一番高くなっている。これは、case2 の報酬の与え方が、エージェントに積極的に行動「recharge」を行なわせていることを示している。

case3 表 5 のような結果となった。

– リサイクルロボットがバッテリーを無駄なく使えるように、バッテリーが減る場合の報酬を行動ごとにそれぞれ減らした。
– 行動「recharge」以外で状態「GOAL」になるということは、他の存在に助けてもらう必要がある。そのため、リサイクルロボット単体で完結するように、行動「recharge」以外で次状態が「GOAL」になる全ての行動の報酬を「-7」にした。
– 上記の理由で、行動「recharge」の報酬を「7」にした。
– リサイクルロボットに積極的に「search」してもらいたいため、行動「wait」の報酬を「-1」にした。

2.2　課題 2

2.3　課題 3

表 6　case3 における方策値と Q 値更新回数の比較

学習率「0.1」, 割引率「0.9」			
状態	行動	方策値	Q 値更新数
high	search	11.194368	54
high	wait	4.146994	54
low	recharge	4.310171	94
low	search	1.989767	94
low	wait	1.20334	94

学習率「0.2」, 割引率「0.9」			
状態	行動	方策値	Q 値更新数
high	search	11.194368	54
high	wait	4.146994	54
low	recharge	4.310171	94
low	search	1.989767	94
low	wait	1.20334	94

学習率「0.1」, 割引率「0.8」			
状態	行動	方策値	Q 値更新数
high	search	11.194368	54
high	wait	4.146994	54
low	recharge	4.310171	94
low	search	1.989767	94
low	wait	1.20334	94

学習率「0.3」, 割引率「0.9」			
状態	行動	方策値	Q 値更新数
high	search	11.194368	54
high	wait	4.146994	54
low	recharge	4.310171	94
low	search	1.989767	94
low	wait	1.20334	94

学習率「0.1」, 割引率「0.7」			
状態	行動	方策値	Q 値更新数
high	search	11.194368	54
high	wait	4.146994	54
low	recharge	4.310171	94
low	search	1.989767	94
low	wait	1.20334	94

学習率「0.4」, 割引率「0.9」			
状態	行動	方策値	Q 値更新数
high	search	11.194368	54
high	wait	4.146994	54
low	recharge	4.310171	94
low	search	1.989767	94
low	wait	1.20334	94

学習率「0.1」, 割引率「0.6」			
状態	行動	方策値	Q 値更新数
high	search	11.194368	54
high	wait	4.146994	54
low	recharge	4.310171	94
low	search	1.989767	94
low	wait	1.20334	94

学習率「0.5」, 割引率「0.9」			
状態	行動	方策値	Q 値更新数
high	search	11.194368	54
high	wait	4.146994	54
low	recharge	4.310171	94
low	search	1.989767	94
low	wait	1.20334	94

学習率「0.1」, 割引率「0.5」			
状態	行動	方策値	Q 値更新数
high	search	11.194368	54
high	wait	4.146994	54
low	recharge	4.310171	94
low	search	1.989767	94
low	wait	1.20334	94

参考文献

[1] https://www.codereading.com/algo_and_ds/algo/quick_sort.html

[2] http://www.ics.kagoshima-u.ac.jp/~fuchida/edu/algorithm/sort-algorithm/quick-sort.html