資源制約下での複数エージェントの協調

Multi-Agent Cooperation under resource bounded environment

○　疋田晃一　(大阪大学)　　高橋泰岳　(大阪大学)　　正　浅田　稔　(大阪大学)

Kouichi HIKITA : Osaka University　　Yasutake TAKAHASHI : Osaka University
Minoru ASADA : Osaka University

Abstract— Robots have to re-charge their own batteries when they try to stay autonomously operational as long as possible. When the number of charging station is less than the number of robots, a competitive situation may happen, and therefore cooperation between the robots is needed to avoid it. In this paper, we investigate how reinforcement learning agents acquire a cooperation behavior under resource bounded environment.

Key Words: reinforcement learning, multi-agent cooperation, resource bounded environment, RoboCup

1.　はじめに

長時間自立的に行動し続けることが要求されるロボットは活動を停止させないようにエネルギーの再補給を行う必要がある．また複数のロボットが存在し，補給を行う場所が1つしか無いような場合にはロボット間で競合が生じることがあり，これを回避するための協調を行わなければならない．

自律的にエネルギーの再補給するロボットを取り上げた研究としてBirkの研究[1]がある．この研究では光センサなどのセンサの出力と停止などの基本的な行動との様々な組合せから，充電ステーションに留まるなどのより効率良く生き続けられる行動を学習により獲得している．この研究は1台のロボットを取り上げたものであるが，複数のロボットを取り上げた研究としてMcFarlandらの研究[2]がある．この研究では固定的行動政策をロボットに埋め込むことにより協調行動を実現しているが，学習は行っていない．資源制約下での複数のロボットの協調を学習により獲得した研究は，筆者らの知る限り見当たらない．しかし未知環境や環境が動的に変化する場合に対処するには学習が有効であると考えられる．

そこで本論文では強化学習を用いてより長い時間生き続ける行動政策を獲得し，資源制約下での複数エージェントの協調や競合の解消がどのように実現されるかについて検証する．

2.　問題設定

環境中に充電ステーションが1台，ロボットが2台存在し，2台はお互いのバッテリ残量の情報を共有している．またロボットは充放電現象に関して先験的な知識を持っていないものとする．本研究で用いるロボットでは充電能力が放電量を下回っているためできる限り2台が長く生き続けることを学習する．

3.　学習の設定

状態変数

- 自分のバッテリ残量:W_{own}
- 他者のバッテリ残量:W_{other}
- 充電ステーションの状態:s

行動は「充電する」,「充電しない」の2通りとし，どちらかのロボットのバッテリ残量が0になったときに負の報酬を与える．行動政策は各ロボットが個々に学習する.両方のロボットが同時に充電しようとした場合，つまり競合が生じた場合は50% の確率でどちらかのロボットの行動を選択する.学習中の行動選択は ε の確率でランダム，それ以外はその時点での最適行動を選択する．

なお実際のロボットでは学習に時間がかかるため学習はシミュレーション上で行い，学習により獲得した行動政策を実際のロボットへ実装する．

4.　学習結果

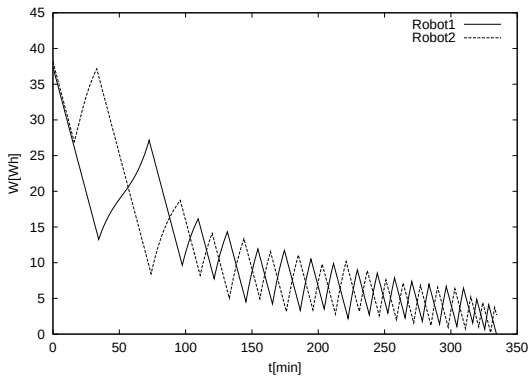


Fig.1 The simulation of the transition of the battery energy

学習により獲得した行動政策を用いて充放電現象をシミュレーションした結果を**Fig.1**に示す.20分あたりで2台のバッテリ残量がほぼ等しいためロボット1の方が早目に充電を行うという協調行動が見られる.また競合に関してはこの図では生じていないが，獲得した行動政策を調べた結果，致命的な競合の回避がなされていることが分かった.致命的な競合というのは例えばロボット1のバッテリ残量が0に近く，ロボット2のバッテリ残量は満充電に近いにもかかわらずロボット2のほうも充電しようとするような場合の競合のことである．

5.　実機への実装

学習により獲得した行動政策の実機への実装を行った.なお本研究ではロボットが自律的にバッテリの充電を行えるようにロボットの回路の改良や充電ステーションの製作も行った．**Fig.2**に製作した充電ステーションと用いたロボットを，**Fig.3**に充電ステーションに入るところを示す.

実機のバッテリ残量の変化を**Fig.4**に示す．**Fig.1**と比較すると充電の周期などに違いが見られるが，これは実際のロボットのセンサ情報にはノイズが重畳しているためシミュレーションとは違う状態へと遷移し，シミュレーションとは違う行動を選択することがあるためだと考えられる．しかし協調的な補給行動が見られ，活動時間が延びていることから学習により獲得した行動政策が実機に対しても有効であると考えらる．

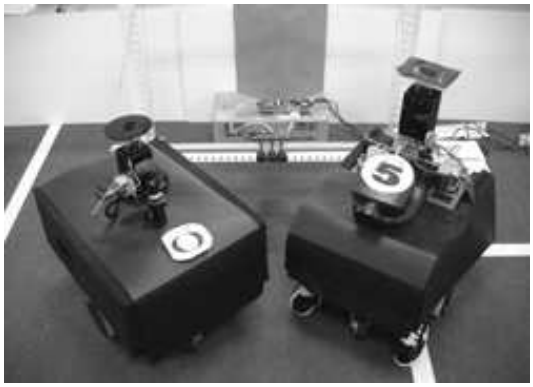


Fig.2 The charging station and robots

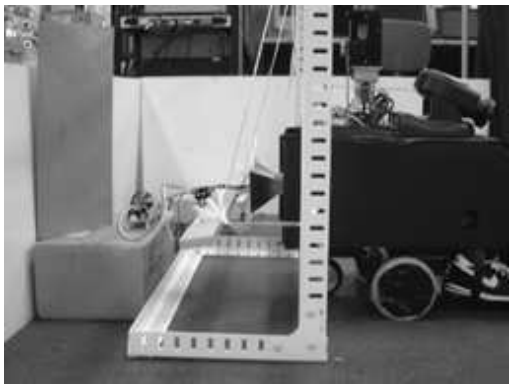


Fig.3 The side view of the charging station when the robot is contacting

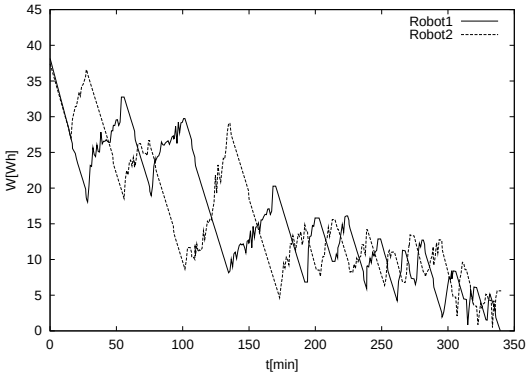


Fig.4 The transition of the battery energy

6. 閾値政策との比較

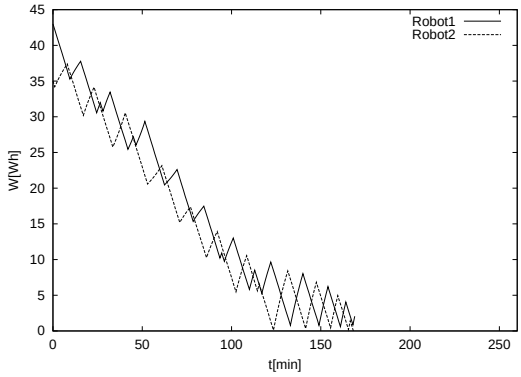
閾値政策として以下のような政策を用いた．

1. $W_{th}, \Delta W_{th}$ を設定
2. $\frac{W_{own}+W_{other}}{2} \geq W_{th}$ のとき $W_{th} = W_{th} - \Delta W_{th}$
3. $W_{own} \geq W_{th}$ のとき充電を開始
4. $W_{other} \geq W_{th}$ のとき充電を終了
5. 2に戻る

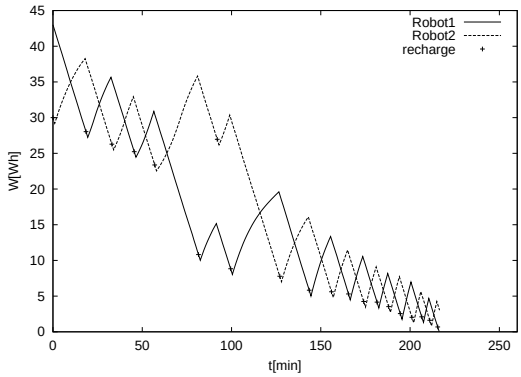
今回は閾値 W_{th} の初期値を $35.0[Wh]$, ΔW_{th} を $5.0[Wh]$ とした．この閾値モデルと消費電力が時々刻々変わるような状況で学習を行い，獲得した行動政策との比較を行う．

まず放電時の消費電力が一定の場合の充放電現象のシミュレーションをこの閾値政策と学習により獲得された行動政策を用いて行った．その結果閾値政策の場合は238分まで，学習により獲得した政策の場合は254分まで生き続けた．

次に行っているタスクや環境の変化により消費電力が時々刻々変わるような状況を想定して充放電現象のシミュレーションをこの2つの政策に対して行った．その結果を**Fig.5**に示す．閾値政策では167分と先程に比べて約70分



(a) Adaptive threshold policy



(b) Learned policy

Fig.5 The comparison between the threshold policy and the learned policy

生き続けた時間が短いのに対し，学習により獲得した政策では219分と先程に比べて35分しか短くなっていない．このことから学習により獲得した行動政策の方が環境の変動に対してロバストであると考えられる．

7. まとめ

強化学習を用いて資源制約下においてより長い時間活動し続ける行動政策を獲得し，協調的な補給行動が実現されていることを示した．今後の課題としては充電ステーションやロボットの数が増えた場合の検証がある．

参考文献

1) A.Birk. Robot Learning and Self-Sufficiency: What the energy-level can tell us about a robot's performance. 6th European Workshop on Learning Robots,August 1-2,1997.
2) McFarland,D. and Bosser,M. Intelligent behavior in animals and robots. MIT Press,1993.