強化学習を用いたサブゴールの発見

室蘭工業大学 情報電子工学系学科 4年 認知ロボティクス研究室 小橋遼

1. はじめに

近年，ロボットが周囲の環境に合わせて自律的に行動を学習する研究およびその実用化が進んでいる．その学習手法の一つとして，強化学習というものが注目されている．

強化学習とは機械学習の一種である．エージェントとも呼ばれる学習者がある状態である行動を取った時に，環境から報酬を受け取る[1]．報酬には行動後すぐに得られる即時報酬と，ある程度行動した後に得られる遅延報酬があり，いつどちらの報酬を与えるかや，報酬の量はユーザー等が設定する．エージェントはより良い報酬を得られる行動を選択していき，最終的に得られる累積報酬が最大になるように学習を進める．強化学習に基づく学習の例としては，スタートからゴールまでの経路探索問題がある(図 1)．ここで，スタートはエージェントの初期位置，ゴールはエージェントが到達すると報酬を得られる位置とする．

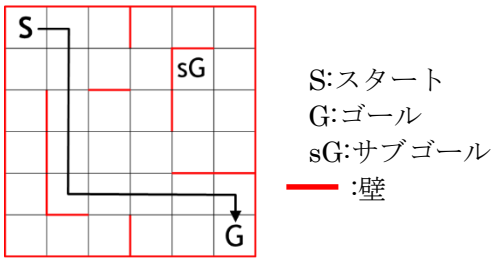

図 1 経路探索問題

今回，エージェントがゴールに到達した時に得られる報酬が普段より大きくなる場合があるタスクを考える．この報酬が大きくなった時に通過している特定の場所をサブゴールと定義する．サブゴールにエージェントが到達しても報酬は得られず，またエージェントはサブゴールの存在や位置を知らされていない．

先述の経路探索問題においてサブゴールが存在する時，強化学習にて学習を行うエージェントがサブゴールを発見できるかどうかを考える．エージェントが認識しているのは現在自分がいる位置だけであり，その位置でどのような行動を選択すればよいかを選択していく．ゴールで報酬を得ることができるので，それを手掛かりにエージェントはゴールの位置を認識している．結果としては，エージェントはゴールまでの最短経路を学習する．この時サブゴールを通過するか否かは，サブゴールが最短経路上に存在するかどうかによる．つまりエージェントはサブゴールの位置を認識していないことになる．これはエージェントがサブゴールに到達しても報酬が得られないので，サブゴールとその他の位置との区別がついていないためである．

強化学習だけではサブゴールを発見できないため，サブゴールを扱う従来の研究では外部からエージェントにサブゴールの位置を知らせるという方法をとっている[2] (図 2)．しかし，この方法では知らせる側がサブゴールの位置やエージェントの現状を知っている必要があり，未知の環境やエージェントの現状を知ることができない状況において学習を行う場合には，サブゴールを発見できない可能性が高い．この問題を解決するために，本研究ではエージェントがサブゴールの自律的発見と，サブゴールを通過する経路の学習を行うことを目指す．

2. 提案手法

今回提案する手法では，エージェントによる自律的なサブゴールの発見と，サブゴールを通過しゴールへ至る行動選択を行い，スタートからサブゴールを経由してゴールへ到達する経路を学習する．

中間発表までは，強化学習に基づく学習ではサブゴールの発見およびサブゴールを通過する経路の学習ができないことを検証するプレ実験を行ってきた．その実験設定を次に示す．

3. プレ実験設定

実験はシミュレーション上で行う．学習を行う環境は3×3のグリッドワールドで，スタート，ゴール，サブゴールがそれぞれ存在する．サブゴールは一つだけ存在し，位置が変化したり消滅することはない．周囲は壁に囲まれていて外に出ることはできず，内部にも一部壁が存在する()．エージェントは上下左右いずれかへ 1 マス移動する行動を選択し，強化学習に基づいてスタートからゴールまでの経路を探索する．学習手法は Q 学習，行動選択手法は ε -greedy 法を用いる．スタートからゴールへ到達するまでを 1 試行とし，これを 200 回行った．実験パラメータは表 1 に示す．

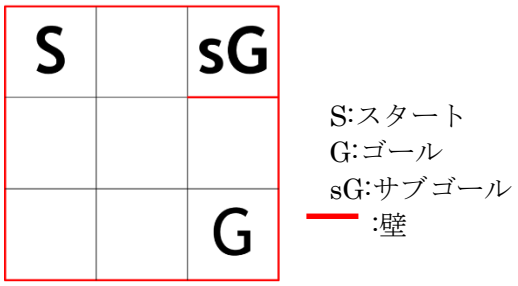

図 2 実験環境

表 1 実験パラメータ

学習率	0.1
割引率	0.9
報酬	100(サブゴール未通過) 1000(サブゴール通過)
ε	0.01

4. 実験結果

この実験の結果を図 3 と図 4 に示す．図 3 は試行毎にゴールで得られた報酬量を，図 4 は試行毎のゴールまでの行動数を表している．図 3 を見ると，サブゴールを通過してゴールへ到達した試行はほとんどないことがわかる．また図 4 を見ると，エージェントの行動数は最終的に 4 へ収束している．この行動数 4 は本実験の環境における最少行動数なので，エージェントはスタートからゴールまでの最短経路を学習している．

これらのことから，エージェントはサブゴールを発見できなかったといえる．なぜこのような結果になったかを考えると，エージェントは報酬を得ることによってその位置がゴールであると認識している一方，サブゴールでは報酬が得られないので，エージェントから見ればサブゴールとその他の場所との区別がつかない．そのため，もしエージェントがサブゴールを通過しても，サブゴールを通過したという認識がないので，エージェントがサブゴールを発見するに至らなかったと考察する．

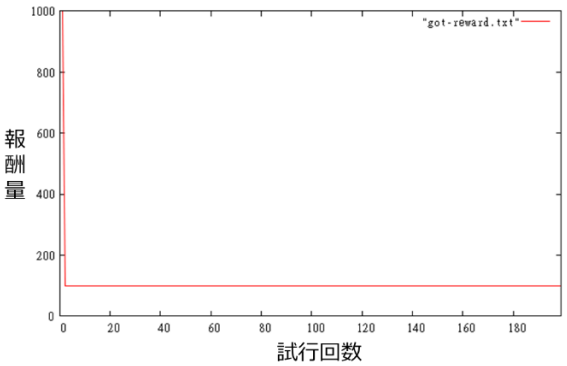

図 3 試行毎の報酬量

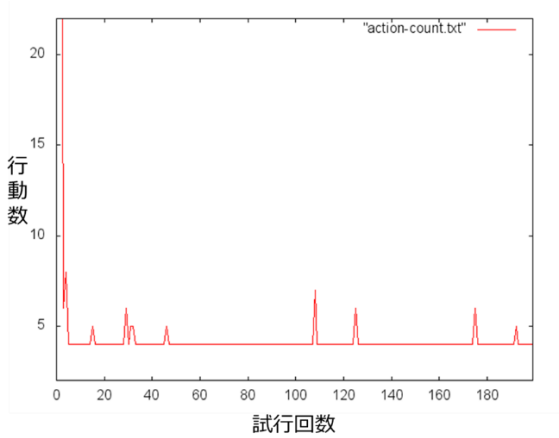

図 4 試行毎の行動数

5. 今後の予定

中間発表以降の予定としては，サブゴールを発見し，かつサブゴールを通過する経路を学習することのできるアルゴリズムの開発を目指す．エージェントがサブゴールを発見するために，ゴールに到達した時に得られる報酬の違いに注目する．報酬の量はサブゴールを通過しているか否かで変わるので，報酬の違いに関わる位置がサブゴールであるといえる．具体的には報酬が小さい時に通っていなくて，報酬が大きい時に通っている場所の内いずれかがサブゴールである．

報酬の違いに関わる位置の見つけ方として，ある試行とその 1 つ前の試行の報酬量と通った経路を比較し，報酬と経路それぞれの違いをエージェントが調べるという方法をとった．しかしこの方法では調べる対象が 2 試行分だけなので，情報量が少なくサブゴールの特定には至らなかった．現在は，ある試行とそれ以前の試行全てを対象に報酬と経路を調べ，サブゴールを発見する方法を考案中である．

サブゴールを発見した後は，通過する経路を学習する．その方法として考えているのは，サブゴールへ向かう行動と，サブゴールからゴールへ向かう行動に分割して学習する方法である．こちらは具体的な方法はまだ決まっていない．

参考文献

[1] 小池康晴，鮫島和行：強化学習の基礎，
[2] 森実克，山田誠二，豊田順一：移動ロボットによるサブゴール間巡回行動の学習，日本ロボット学会誌，P1001～1004(1994 年)