報酬の差異による単体サブゴール発見手法の提案

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

近年では，ロボットが自律的に行動を学習する研究が進んでおり，中でもロボットに報酬を与えることでこれを可能とする強化学習が注目されている．学習を行うロボットがタスクを実行するにあたって，タスクを達成するまでの小目標が存在する場合がある．この小目標をサブゴールと定義する．

サブゴールが存在するタスクを扱う従来研究では，人間がサブゴールのクリア時にロボットへ報酬を与えることで，サブゴールをクリアする行動の学習を可能としている[1]．

従来研究では，サブゴールでロボットに報酬を与えることで，サブゴールをクリアする行動の学習を可能としている．これはロボットが使用される環境を事前に想定できるからである．学習ロボットは周りの環境に対して自律的に行動を学習可能なので未知の環境で用いられることが多い．しかし，未知の環境ではロボットがクリアすべきサブゴールが想定できない可能性がある．従って，あらゆる状況に応じて人間がサブゴールで与える報酬を設定することは難しいという問題がある．

本研究では，あらゆる環境においてロボットがサブゴールをクリアする行動の学習を実現するために，サブゴールを自律的に発見しサブゴールをクリアする行動の学習を行う学習システムを提案する．これにより，サブゴールの発見およびサブゴールをクリアする行動を自律的に行うロボットの実現を目標とする．

5. アプローチ

本研究では，経験するとタスク達成時に与えられる報酬が大きくなる特定の状態が存在する時，その状態をサブゴールとして扱う．与えられる報酬の差異からサブゴールを発見し，発見したサブゴールをクリアする行動を学習させる．

6. 提案システム

まずサブゴールの探索を行う時の提案システムの動作を図 1 に示す．

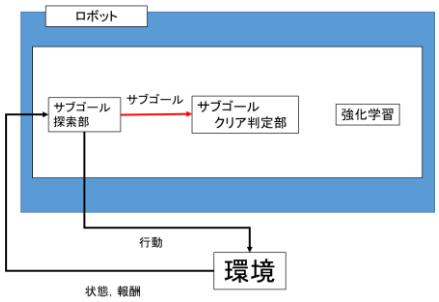

図 1 提案システム概要：サブゴール探索時

サブゴール探索部では，1 試行中にロボットが認識し，経験した状態を蓄積する．その後自身が取り得る行動の中からランダムに行動を選択し，行動の結果に応じて報酬r_tを与えられる．試行終了後，試行中に経験した状態をr_tの値に応じて分類し蓄積する．ある試行 t において与えられる報酬r_tを式(1)のように定義する．

$$r_t := \begin{cases} r_{small}(\text{サブゴールをクリアしていない}) \\ r_{big}(\text{サブゴールをクリアしている}) \end{cases} (r_{small} < r_{big}) \quad (1)$$

ロボットが t 試行目に経験した状態の集合を$\boldsymbol{E}_t$とすると，$r_t = r_{big}$の時常に経験した状態の集合$\boldsymbol{E}_{big}$と，$r_t = r_{small}$の時一度でも経験したことのある状態の集合$\boldsymbol{E}_{small}$としてそれぞれ蓄積する．$\boldsymbol{E}_{big}$，$\boldsymbol{E}_{small}$それぞれを式(2)，式(3)で定義する．

$$\boldsymbol{E}_{big} := \bigcap_{t \in T_{big}} \boldsymbol{E}_t \quad (2)$$

$$T_{big} := \{t | t \in \boldsymbol{N} \wedge r_t = r_{big}\}$$

$\boldsymbol{N}$：自然数の集合

$$\boldsymbol{E}_{small} := \bigcup_{t \in T_{small}} \boldsymbol{E}_t \quad (3)$$

$$T_{small} := \{t | t \in \boldsymbol{N} \wedge r_t = r_{small}\}$$

$\boldsymbol{N}$：自然数の集合

特定の試行終了後サブゴールの探索を終了し，$\boldsymbol{E}_{big}$および$\boldsymbol{E}_{small}$からサブゴール sG を割り出す．サブゴールとなり得る状態は，報酬が小さい試行では一度も経験しておらず，報酬が大きい試行では必ず経験している状態である．このことから，サブゴール sG を式(4)のように求め，発見したサブゴールをサブゴールクリア判定部へ渡す．

$$\mathrm{sG} := \boldsymbol{E}_{big} \cap \overline{\boldsymbol{E}_{small}} \quad (4)$$

サブゴールを発見した後は，発見したサブゴールをクリアしてからタスクを達成する行動を学習する．

次にサブゴールをクリアする行動を学習する時の動作を図 2 に示す.

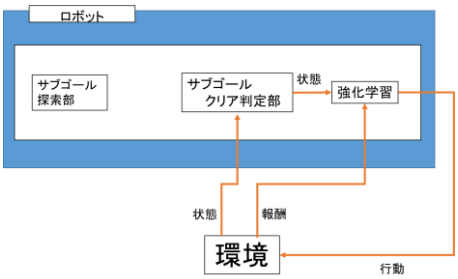

図 2 提案システム概要：行動学習時

まずロボットは現在の自身の状態を認識する. 次に認識した状態が発見したサブゴールと一致するか否かを，サブゴールクリア判定部にて判定する. 認識した状態がサブゴールでなければ通常の強化学習を行うが，サブゴールと一致していれば，図 3 のように状態の追加を行い，その状態を基に強化学習を進める.

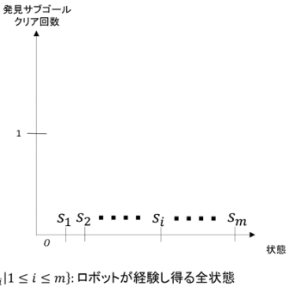

図 3 状態追加後の全状態

7. 実験

実験はシミュレーションで行う. 実験によって提案するシステムを適用したロボットがサブゴールを発見し，発見したサブゴールをクリアする行動が学習できるかどうか検証する. 実験環境は図 4 のような 3×3 のグリッドワールドで, 周囲と内部に壁が存在する. スタート，ゴール，サブゴールが存在し，ロボットはゴールに到達すると報酬を得られる. またサブゴールを通過してゴールへ到達すると，より多くの報酬を得ることができる. 学習を行うロボットは強化学習によって行動学習を行うロボットＡと，提案システムを適用したロボットＢの二体を設定する. どちらのロボットも選択可能な行動は共通で，上下左右への移動が可能であり，壁に衝突した時はその場に停止する.

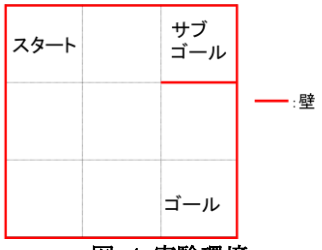

図 4 実験環境

本実験のパラメータを表 1 に示す.

表 1 実験パラメータ

学習率	0.10
割引率	0.90
報酬	100.0（サブゴール未通過） 1000.0（サブゴール通過）
試行回数	20000 回(行動学習時) 100 回(サブゴール探索時)
ε (ε – greedy 法)	0.05
サブゴールの数	1

8. 実験結果と考察

図 5 にロボットＡ，Ｂそれぞれが試行毎に与えられた報酬を，図 6 に 19 万試行から 20 万試行に与えられた報酬を示す.

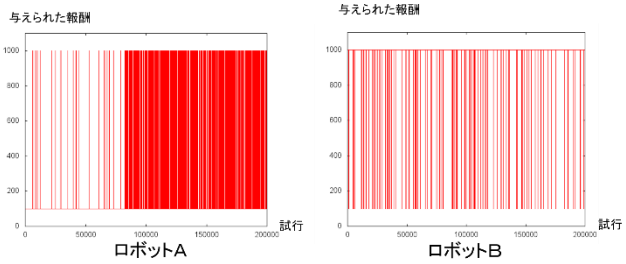

図 5 各ロボットに与えられた報酬

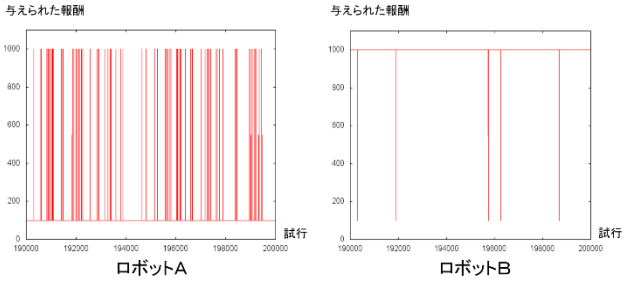

図 6 各ロボットに与えられた報酬(19 万～20 万試行)

実験結果より，ロボットＡは試行を重ねても 1000 の報酬を得る行動を学習することができず，サブゴールを通過する行動は学習できていないと考えられる. またロボットＢはほとんどの試行で 1000 の報酬を与えられており, サブゴールを通過する行動を学習できていると考えられる. 以上のことから, サブゴールの発見および状態の追加によってサブゴールを通過する行動の学習が可能となると考えられる.

9. まとめ

シミュレーション実験により，与えられる報酬の差異からサブゴールを発見し，クリアする行動を学習できることが示された.

参考文献

[1] 前田　陽一郎, 花香　敏“Shaping 強化学習を用いた自律エージェントの行動獲得支援手法”, 日本知能ファジィ学会誌, Vol.21, No.5, pp.722-733（2009）