強化学習における報酬非依存型知識の利用

宮崎　愛央
室蘭工業大学　情報工学科　4年　認知ロボティクス研究室

Abstract

現在，機械学習と呼ばれるものはいくつかあり，強化学習もその1つである．本研究では強化学習において報酬に依存しない知識というものに注目し，それを用いることで同環境下において様々なタスクに対応させることを目的とする．

1　はじめに

1-1　背景

近年，家庭用のロボットなど我々の生活環境と同等の環境で働くようなロボットが見られるようになった．しかし，様々な環境で働くことからロボットは様々なタスクをこなす必要が出てきた．そのために機械学習と呼ばれる手法によってロボット自ら学習を行わせることに注目が置かれた．機械学習の中でも特に強化学習と呼ばれる手法は実ロボットで用いられることが多い手法として注目されている．

1-2　強化学習概要

強化学習はある状態で取った行動に対して報酬を得ることで学習を行う手法である．この報酬はロボットが目的を果たせるように人間が設定する．

1-3　強化学習の問題点

強化学習では報酬によって学習が行われるが，果たすべき目的が変わってしまった場合に対応が遅れてしまうといった問題点がある．

2　報酬非依存型知識の提案

2-1　報酬非依存型知識の定義

報酬に依存しない情報として，ロボット（エージェント）から見た環境の遷移情報を「報酬非依存型知識」とし，以下のように定義した．

（状態，行動）→（次状態）

また，ロボットは定義した知識を保持するために知識テーブルを持ち，一つ一つの報酬非依存型知識を独立に扱う．

2-2　提案システム概念図

今回提案するシステムの概要図を Fig.1 に示す．

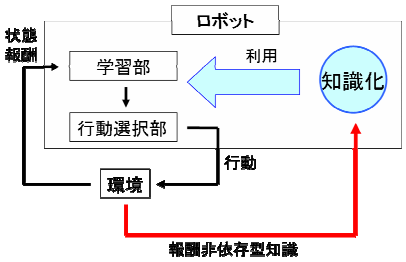

Fig.1:概要図

2-3　提案システムの流れ

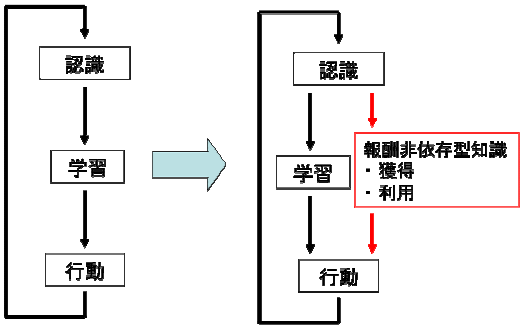

Fig.2:報酬非依存型知識を含む流れ（左が従来の強化学習学習，右が今回提案するシステム）

2-4　ロボットによる報酬非依存型知識の獲得

ロボットは行動毎に自身が持つ知識テーブルの中にその行動による状態遷移情報を追加する．

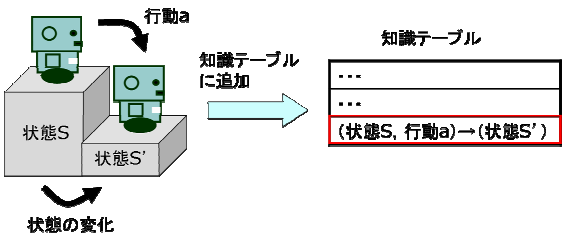

Fig.3:報酬非依存型知識の獲得

2-5　報酬非依存型知識の利用

報酬非依存型知識の利用は目的に対してどのように学習していけば良いかを予測し，予測した行動を取りやすくなるように強化学習で用いる価値関数を更新する．

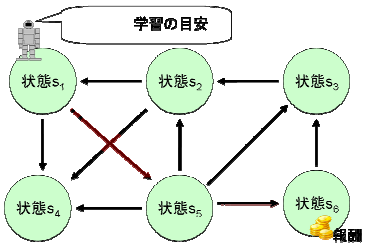

Fig.3:報酬非依存型知識の利用の仕方

2-6　報酬非依存型知識による環境変化への対応

自分の行動の結果と知識テーブルにある情報を比べることで環境変動を認識・対応を行う．

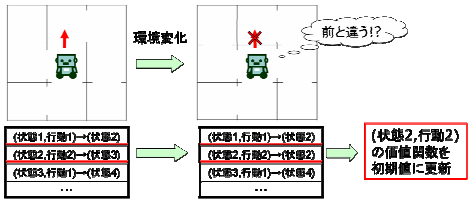

Fig.4:環境変動への対応

3　提案システムを用いた実験

3-1　実験の種類

実験は大きく分けて静的環境下での実験と，動的環境下での実験になる．本稿では静的環境下で行われた実験の 1 つを載せる．

3-2　実験目的

・タスク変化に対する対応が早いことを示す

3-3　実験設定

実験は迷路問題を用いた．今回ゴールは一定回数ゴールするごとに別の場所へ移る．これはタスクの変化を表す設定である．強化学習の手法として Q 学習を用い，行動選択手法として ε -greedy を用いる．また，実験結果を比較するため以下のエージェントで実験を行った．

- A.　報酬非依存型知識を利用しない
- B.　報酬非依存型知識を利用（利用度が低い）
- C.　報酬非依存型知識を利用（利用度が高い）

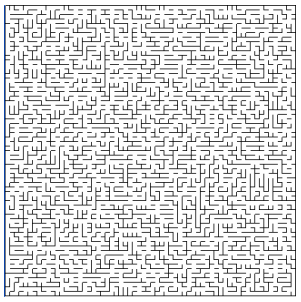

Fig.5:実験で用いた迷路

Table.1:詳細設定

迷路サイズ	64×64
報酬(ゴールのみ)	100
実験終了までの試行回数	1000
ゴールが変化する試行回数	200
Q 値初期値	0.001
α	0.5
γ	0.8
ε	0.05

3-4　結果

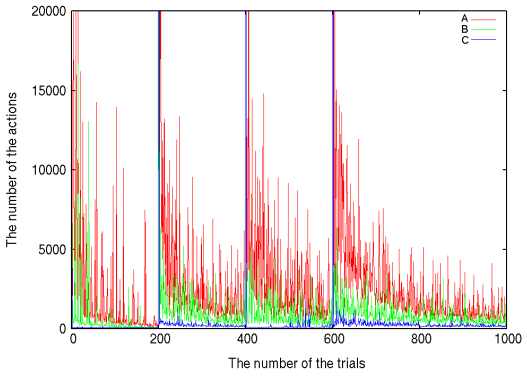

Fig.6:各試行における行動数の比較

3-5　考察・まとめ

Fig.6 は X 軸に試行回数，Y 軸に行動数を取っている．どのエージェントも 200 試行ごとに行動数が急激に増えている．これはゴールの位置が変化した影響を受けていることを表している．しかし，報酬非依存型知識を利用した場合は利用しない場合に比べて各試行における行動数が少なくなっており，ゴール変化の影響を抑えていることがわかる．この実験から報酬非依存型知識の利用はタスク変化に対して有効であるといえる．