確率表現を用いた報酬非依存型知識の提案

澁谷　和
室蘭工業大学　情報工学科　4 年　認知ロボティクス研究室

1　はじめに

1-1　研究背景

近年，ロボットの発達とともに機械学習の必要性が向上してきた．機械学習の中でも特に強化学習[1]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．

強化学習では報酬のみ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．

そこで強化学習の問題点を解決するために宮崎による「強化学習における報酬非依存型知識の利用」[2]という研究がある．報酬非依存型知識という環境遷移に関する情報(報酬に依存しない情報)を知識として活用することにより，目的が変更されても対応が可能となった．しかし，同時に動的環境下では学習効率が低下するという問題が起こった．

この問題を解決するために，本研究では確定的な知識であった報酬非依存型知識を確率的な知識に拡張することを提案し，動的環境への適応を目指す．ここでは，動的環境というのは環境変化が起きる環境のことを指す．

2　先行研究

報酬非依存型知識は環境遷移に関する情報を定義したものである．具体的には，エージェントが認識する状態と行動，行動の結果の遷移先によって構成されている．この状態行動対と遷移先は 1 対 1 対応である．そのため，動的環境において遷移先が変化した場合に対応できない恐れがある．

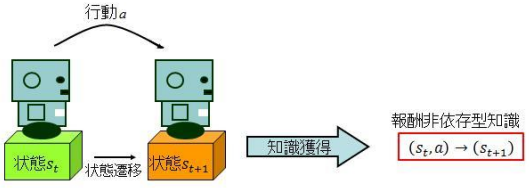

図 1　報酬非依存型知識の例

3　確率的報酬非依存型知識の提案

3-1　確率的報酬非依存型知識の概要

先行研究では状態行動対と遷移先が一対一対応であった．そこで，報酬非依存型知識に複数の遷移先を持たせ，各遷移先に遷移確率を持たせるこの確率化した知識を「確率的報酬非依存型知識」と定義する．

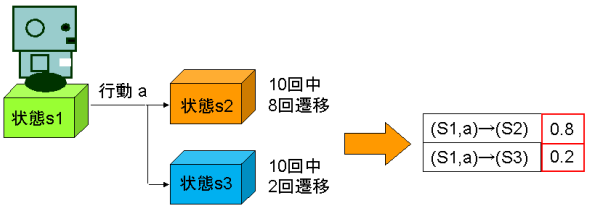

図 2　確率的報酬非依存型知識の例

3-2　確率的報酬非依存型知識の定義

確率的報酬非依存型知識の遷移確率を式(1)で定義する．

$$P_{ss'}^{a} = \Pr\{s_{t+1} = s' \mid s_t = s, a_t = a\} \qquad (1)$$

s_t はエージェントが認識している状態を表す．a_t はエージェントが取る行動を表す．s は状態 s_t において行動 a_t を取った時の遷移先の状態を表す．s は任意の状態を表す．a は任意の行動を表す．s' は任意の次状態を表す．

確率的報酬非依存型知識を式(2)に示す．ここではエージェントの取りうる状態の集合を $S = \{s_1, s_2, s_3, \cdots, s_n\}$ とする．

$$k_{sa} = \{P_{ss'}^{a} \mid s_t = s, a_t = a, s_{t+1} = s', s' \in S\} \qquad (2)$$

エージェントは定義した知識を保持するために知識テーブルを持つ．知識テーブルは式(3)で定義する．エージェントの取りうる行動の集合を $A = \{a_1, a_2, a_3, \cdots, a_n\}$ とする

$$K := \{k_{sa} \mid s \in S, a \in A\} \quad (3)$$

3-3　確率的報酬非依存型知識の獲得

エージェントは行動毎に確率的報酬非依存型知識を獲得し，遷移確率を計算する．

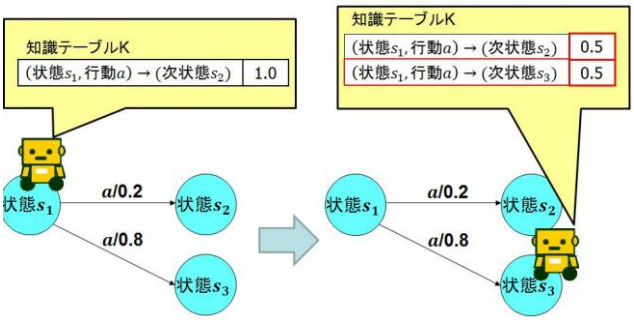

図 3 確率的報酬非依存型知識の獲得

3-4　確率的報酬非依存型知識の利用

確率的報酬非依存型知識を用いて，どのように行動すれば目的状態に辿りつくかの予測を行う．ここでは遷移確率に応じて強化学習の価値関数を更新する．

4　実験

4-1　実験の目的

確率的報酬非依存型知識を用いることにより動的環境に適応できることを示す．

4-2　実験設定

実験は迷路問題を用いた．動的環境を生み出すため，ランダムに動く障害物を迷路に投入する．また，ゴールは一定回数ゴールするごとに別の場所に移る．強化学習の学習手法として Q 学習を用い，行動選択手法として追跡手法を用いる．実験結果を比較するため以下のロボットで実験を行った．

- A.　強化学習のみ
- B.　強化学習+報酬非依存型知識(先行研究)
- C.　強化学習+確率的報酬非依存型知識

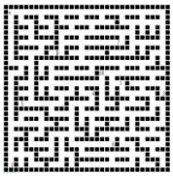

図 4：実験で用いた迷路

表 1　実験パラメータ

迷路サイズ	33×33
障害物の数	40
報酬(ゴールのみ)	100
実験終了までの試行数	2500
ゴールが変化するまでの試行数	250
Q 値の初期値	0.001
学習率 α	0.5
割引率 γ	0.7
β	0.7
知識の利用度 f	1.0

1 行動あたりの障害物との遭遇確率　0.07

4-3　結果

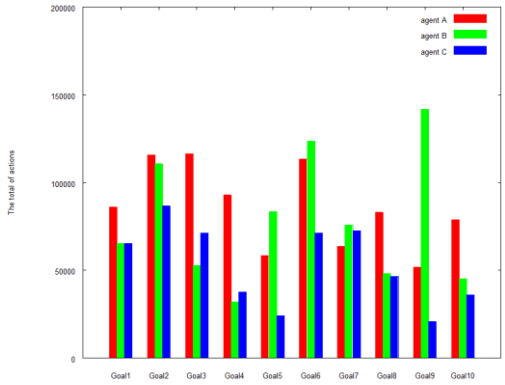

図 2：各ゴールにおける行動総数

5　考察・まとめ

確率的報酬非依存型知識を用いた場合，先行研究の知識よりも行動数が少なくなっている．これは，動的環境下において学習収束が早いためである．この実験により確率的報酬非依存型知識は動的環境下に対応できているといえる．

参考文献

[1] Richard S. Sutton and Andrew G. Barto., “Reinforcement Learning”, The MIT Press, 1998
[2] 宮崎愛央,「強化学習における報酬非依存型知識の利用」, 室蘭工業大学　平成 21 年度卒業研究