センサの重要度に応じた学習空間の動的構成

室蘭工業大学 情報電子工学科 4年 認知ロボティクス研究室 木村 敏久

1. はじめに

近年ロボットに適用される学習手法として強化学習[1]が注目されている．強化学習とは学習者自身が試行錯誤を繰り返す事により，環境に適した行動を学習する学習手法である．

2. 強化学習の問題点

強化学習の学習空間は行動軸と状態軸(センサ軸)によって構成される．そのため，搭載されるセンサが追加されると，対応する状態軸が追加され状態数が増加し学習時間が増加する問題がある．

学習空間が拡大する原因として，学習空間の構成方法に問題がある．複数のタスクを行う場合，タスク達成に対して各センサの重要度は異なると考えられる(図 2.2)．本来であればタスク毎に，センサの重要度に応じてセンサ軸の重みを変えた学習空間を構成し学習を行うのが望ましい．しかし強化学習では，センサの重要度を考慮していない学習空間を構成し学習を行う．そのため状態数が増え学習時間が増加してしまう．センサの重要度に応じて，センサ軸の重みを変えた学習空間を構成できていない事が学習時間増加の原因になっている．

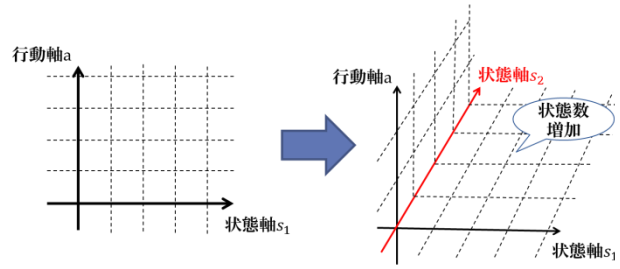

図 2.1　学習空間の拡大

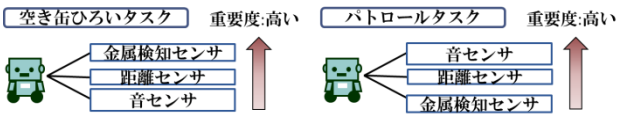

図 2.2　センサの重要度が異なる例

3. 本研究の目的

本研究はロボットが，タスク達成における各センサの重要度に応じて，学習空間を自律的に構成する手法を提案し，強化学習の学習速度を向上させる事を目的とする．

4. アプローチ

センサの重要度を算出するために，センサの出力値と報酬の関係に注目する．例として，高度センサを搭載したロボットが登山タスクを行う事を考える．この時，高度センサの値が高ければ高い程，登山タスクの進捗度が上がるため，高い報酬を受け取れる．このように，センサと報酬には相関関係があると考えられる．この相関関係を利用してセンサの重要度を算出する．

算出したセンサの重要度が高ければ，タスク達成に重要なセンサと予測されるため，学習空間の状態数を多くする．逆に重要度が低ければ，タスク達成に重要ではないセンサと予測されるため，状態数を少なくする(図 3.1)．重要度が低いセンサの状態数が減少するため，状態空間が縮小し学習速度が向上すると考える．

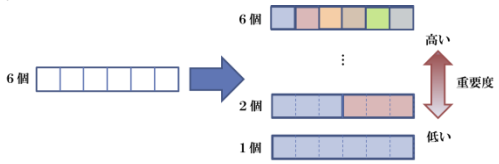

図 3.1　状態数の変更の例

5. 提案手法

提案手法の概念図を図4.1に示す．提案手法は，各センサの重要度を算出する「重要度算出部」，センサの重要度に応じてセンサの状態数を変更する「重要度利用部」によって構成されている．

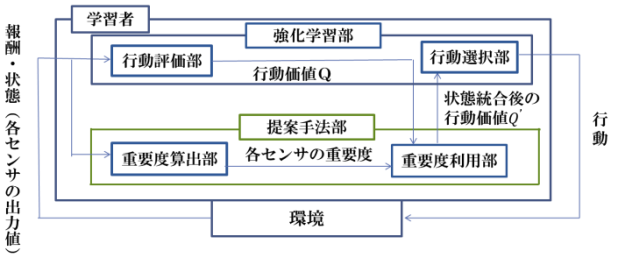

図 4.1　提案手法の概念図

5.1. 重要度算出部

重要度算出部は各センサの重要度を算出する部分である．報酬と状態(各センサの出力値)を毎時刻取得し，センサの重要度を算出するデータとして記憶する．報酬r_tを目的変数，搭載されているp個のセンサの出力値$e_{i,t}(1 \le i \le p)$を説明変数とした，重回帰式(1)により重回帰分析を行う．

$$r_t = e_{1,t}a_1 + e_{2,t}a_2 + \cdots + e_{p,t}a_p + a_0 \qquad (1)$$

tは時刻，$a_i(1 \le i \le p)$は回帰係数，a_0は定数を表している．回帰係数a_iは，センサの出力値$e_{i,t}$の報酬r_tに対しての影響度を表している．よって本研究では，回帰係数a_iを求め，各センサの重要度として利用する．

5.2. 重要度利用部

重要度利用部は各センサの重要度に応じて状態数を変更する部分である．入力として各センサの重要度と行動価値 Q を取得する．各センサの状態数v_iを，センサの重要度a_iに応じて式(2)から算出する．

$$v_i = \begin{cases} v_{i,min} & (|a_i| < m_\alpha) \\ \left\lfloor \frac{v_{i,max}-v_{i,min}}{m_\alpha - m_\beta}|a_i| + \frac{m_\beta v_{i,min} - m_\alpha v_{i,max}}{m_\beta - m_\alpha} \right\rfloor & (m_\alpha < |a_i| < m_\beta) \\ v_{i,max} & (m_{\beta<}|a_i|) \end{cases} \quad (2)$$

$v_{i,min}$はセンサ i が表現できる最小の状態数，$v_{i,max}$はセンサ i が表現できる最大の状態数，m_αとm_βは定数を表している．算出された状態数v_iから，各センサの状態数を変更した Q 空間を一時的に作成する．作成した一時的な Q 空間における現状態s^*の行動a_iに対しての行動価値$\mathrm{Q}(s^*, a_i)$を式(3)から算出する．

$$Q^*(s^*, a_i) = \frac{\sum_{c_k}\sum_{c_l}\cdots\sum_{c_m} Q(s_w, a_i)\cdot N(s_w, a_i)}{\sum_{c_k}\sum_{c_l}\cdots\sum_{c_m} N(s_w, a_i)} \quad (3)$$

ここで，s_wは統合された状態に含まれる元々の状態であり，$\mathrm{Q}(s_w, a_i)$はs_wの行動a_iの行動価値，$\mathrm{N}(s_w, a_i)$は$\mathrm{Q}(s_w, a_i)$の評価回数を表している．

6. 実験

6.1. 実験概要

提案手法と一般的な強化学習の性能を比較するため実機実験を行った．使用するロボットと実験環境を図 6.1 に示す．実験タスクとして前方の「壁 A の近傍に到達する」というタスクを行う(図 6.2)．ロボットは即時報酬として式(4)を行動毎に受け取る．実験パラメータを表 6.1 に示す．d_Aは$0 \le d_A \le 10$の範囲の値を取る．

$$\mathrm{r} = 11 - d_A \quad (4)$$

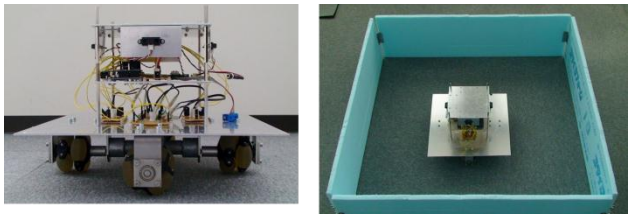

図 6.1 ロボットと実験環境

表 6.1 実験パラメータ

行動学習手法		加重平均手法
行動選択手法		ε -greedy 法
総試行回数		100 回
探査的な行動 ε		0.1
学習率 α		0.5
行動価値 Qの初期値		0.0
最少の状態数$v_{i,min}$		1
最大の状態数$v_{i,max}$		11
定数	m_α	0.2
	m_β	0.8

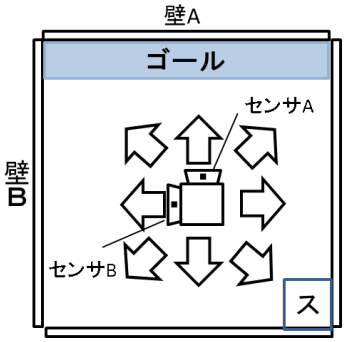

図 6.2　実験タスク

6.2. 実験結果

結果を図 6.3，図 6.4，図 6.5 に示す．図 6.3，図 6.4 からセンサ A の重要度は 1 試行目の途中から「0.85～0.9」に，センサ B の重要度は「0.1～0.15」に収束しているのがわかる．また，重要度に応じてセンサ A の状態数が多く，センサ B の状態数が少なくなっている．図 6.5 から，提案手法が強化学習より早く行動回数が収束しており，学習速度が向上しているのがわかる． よって提案手法は有効的に機能したと考える．

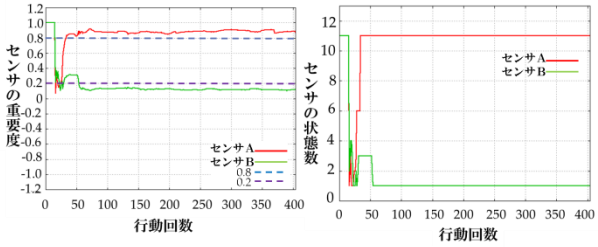

図 6.3　1 試行目のセンサの重要度と状態数

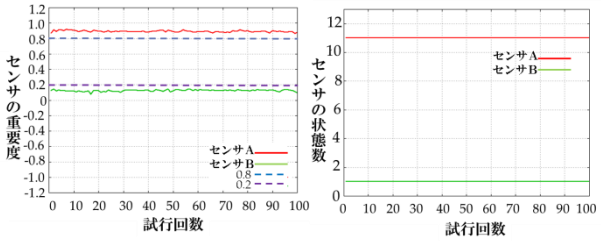

図 6.4 各試行終了時のセンサの重要度と状態数

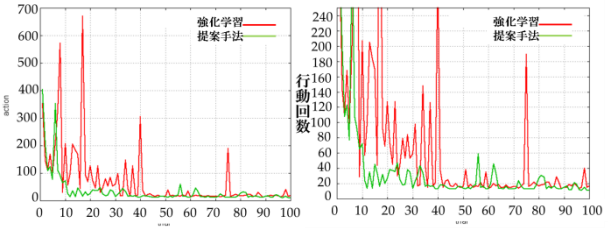

図 6.5　各試行における行動回数の推移

今後の課題として，遅延報酬環境への適用とさらなる検証実験を行う事が挙げられる．

参考文献

[1] Richard S. Sutton, Andrew G. Barto(共訳:三上貞芳，皆川雅章):強化学習，2001 年8 月10 日，第1版第2 版発行