Actor-Critic 法を利用した分人モデルに基づく行動選択

情報科学科　永見 俊樹　　指導教員：小林 邦和, 鈴木 拓央

1　はじめに

今日におけるロボットは、様々な分野で活躍が見られる。特に人とコミュニケーションをとることによって、生活を支援しているロボットが多く存在している。しかしロボットは、人間に対してコミュニケーションをとる際に画一的な対応しか取れないという問題を抱えている。それに対して、人間同士でのコミュニケーションでは相手に応じて様々な対応をとることができる。

本研究では分人モデルを利用してロボットがどの人間とインタラクションをしているのかを選択、また様々な行動の中からそれぞれの人にあった行動を選択するということを実現する。

2　分人

分人とは、対面する人間によって接し方が変化するという概念であり、小説家の平野啓一郎によって提唱された [1]。コミュニケーションをとるたびに人の頭の中では相手の分人が形成され、その後の関わり方によって対象の分人が成長していく。その分人にはパターンが 3 つあり図 1 のように社会的な分人、グループ向けの分人、特定の相手に向けられた分人と分けられる。

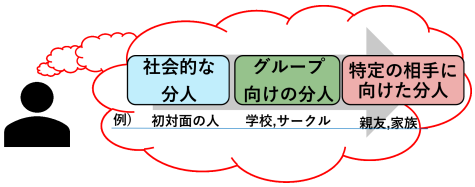

図 1　分人の種類をを表す概念図

3　分人モデル

本研究で利用しているモデルは、先行研究である田中らの分人モデルである [2]。分人モデルの構造としては図 2 のようであり、入力として個人の情報、出力にそれぞれのモジュールの状態価値、行動価値をとっている。また分人という概念を実現するためにモジュール型ニューラルネットワークを導入している。これは 1 つのモジュールを 1 つの分人に対応させるためである。またそれぞれのモジュールを学習させるために、図 3 のように Actor-Critic 法をニューラルネットワークを用いて表現している [3]。

先行研究では行動の出力に関して分人のパターン（図 1）を出力しており分人の成長が可視化されていた。本研究ではインタラクションをしている相手がどの分人に対応するかを評価した後、その評価した相手に対して適切な行動を出力するように出力の変更を加えている。

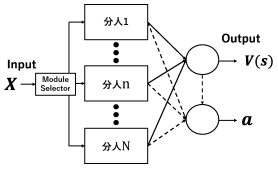

図 2　分人モデル

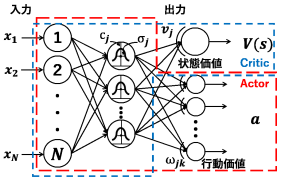

図 3　Actor-Critic の構造

またモデルのパラメータ更新式、更新に必要な TD 誤差、モジュールの選択確率は、それぞれ以下の式で表される。

TD 誤差　critic は報酬 r と次の状態 $\boldsymbol{s}' = (x_1, x_2, \cdots, x_N)^T$ を観測し、critic と actor の学習指標となる TD 誤差 (temporal difference error) を計算する。ある時刻での TD 誤差 δ は、式 (1) で表される。

$$\delta = r + \gamma V(\boldsymbol{s}') - V(\boldsymbol{s}) \quad (0 \leq \gamma \leq 1) \tag{1}$$

中間ユニットの学習　中間層の出力関数は式 (2) に示すガウス関数 y_j を用いる。平均を $\boldsymbol{c}_j$、分散を σ_j^2 とし、式 (3) で学習を行う。

$$y_j = \exp\left(-\frac{\|\boldsymbol{s} - \boldsymbol{c}_j\|^2}{2\sigma_j^2}\right) \tag{2}$$

$$\boldsymbol{c}_j \leftarrow \boldsymbol{c}_j + \zeta\,\delta\,\nu_j\,\frac{\boldsymbol{s} - \boldsymbol{c}_j}{\sigma_j^2}\,y_j \quad (0 \leq \zeta \leq 1) \tag{3}$$

critic の学習　TD 誤差を零に近づけるように、式 (4) で重み ν_j の学習を行う。

$$\nu_j \leftarrow \nu_j + \eta\,\delta\,y_j \quad (0 \leq \eta \leq 1) \tag{4}$$

actor の学習　状態価値 $V(\boldsymbol{s})$ が高くなるように、式 (5) で重み ω_j の学習を行う。

$$\omega_{jk} \leftarrow \omega_{jk} - \rho\,(f(a_k) - t_k)\,f'(a_k)\,y_j \tag{5}$$

モジュールの選択確率　式 (6) と式 (7) の Gibbs 分布によるソフトマックス手法により、設計された入力情報が適切なモジュールへ入力されるように学習を行う。

$$\pi(\boldsymbol{s}) = \frac{\exp(p(\boldsymbol{s}))}{\sum_{\boldsymbol{s}' \in X} \exp(p(\boldsymbol{s}'))} \tag{6}$$

$$p(\boldsymbol{s}) \leftarrow p(\boldsymbol{s}) + \beta\,\delta \quad (0 \leq \beta \leq 1) \tag{7}$$

4　計算機シミュレーション

入力の個人情報は 2 つに分かれており 1 つ目は名前、国籍、性別のような個人を特定する情報である「カテゴリー情報」、2 つ目は状態、その状態でとり得る行動のパターンなどが含まれている「行動情報」である。本研究ではロボットが飲食店の店員、お客さんが来店する際に状態として {前菜、メイン、デザート} という 3 つの状態をとりそれぞれの状態で 9 つの商品を提供する。「行動情報」には状態とその状態に対する良い行動を設定する。

図 4 では学習後に A,B,C の 3 人が順に来店するという場面設定で適切な分人を選択できているかどうかを縦軸が分人のモジュールの状態価値、横軸インタラクションの回数としてが選択している。高く反応があるモジュールが選択され、分人の選択を行っている。図 5 では A さんが選択された際、9 つの行動のうちどの行動が選択されるかを縦軸が行動価値、横軸が行動パターンとして選択している。A さんは入力として行動 {1,4,7} を適切な行動と設定しているため確実に選びたい行動の価値が高まっていることがわかる。

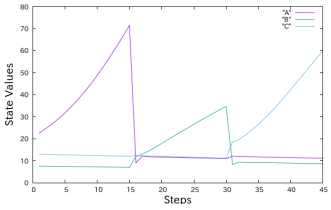

図 4　各モジュールの出力

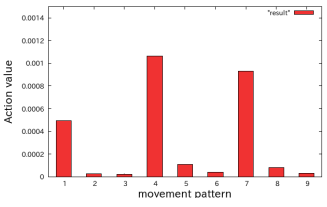

図 5　行動価値の出力

5　おわりに

本研究では、先行研究における抽象的な行動出力の具体化という課題の解決に務めた。分人が形成されつつ実際に行動が出力されることを確認した。今後の課題としては、Actor-Critic 法の強みである連続値を扱うことで行動選択をすること、実際にロボットへの実装ということが挙げられる。

参考文献

[1] 平野 啓一郎, 「私とは何か「個人」から「分人」へ」, 講談社新書 2012

[2] 田中 利幸, 小林 邦和 「相互適応的 HRI を思考した学習型分人モデルの構築に関する研究」, 愛知県立大学, 大学院情報科学研究科修士論文, 2016

[3] 三上 貞芳, 皆川 雅章, 「強化学習」, 森北出版株式会社,2000