深層強化学習モデルの内部挙動の言語化による制御手法構築へ向けて

圓田彩乃　（指導教員：小林一郎）

1　はじめに

近年、様々な場面で深層学習が用いられているが，一般に深層学習モデルの内部挙動はブラックボックスであることから，用途によっては使用が制限される．そのような背景を踏まえ，本研究では深層学習モデルの内部挙動が人間が理解できるように言葉で説明することを目指す．アプローチ方法として，深層学習モデルで得られた入出力関係をファジィモデリングし、その関係をファジィ言語変数からなる規則で表現することにより，モデルの入出力の振る舞いを人間が把握しやすいようにする．具体的には，CartPole を題材として深層強化学習によって学習されたモデルの制御規則を言語で説明する．また，その制御規則を用いて CartPole を制御することで，生成された制御規則の正当性を示す．

2　深層強化学習を用いた CartPole の制御

深層学習モデルの内部挙動の言語化を対象として，強化学習の手法の一つである Q 学習を深層にした Deep Q Network(DQN) [1] を用いて CartPole の学習を進める．CartPole とは，台車を右か左に押すことで台車の上の棒が倒れないように制御するタスクである．本研究では，1 step につき 1 回 action を取り，「棒が 12° 以上倒れる」「台車が画面から出る」「200 step 到達」のいずれかを満たすと 1 エピソードが終章し，報酬は 1 step 倒立していられるごとに 1，さらに 195 step まで到達すると追加報酬 1 が与えられるものを採用する．CartPole への入力は台車の位置，台車の速度，棒の角度，棒の角速度の 4 種類，出力はその時にとった action の 1 種類である．action は 0〜20 の力で右または左に押す 41 通りある．今回は，直近 10 エピソードの平均報酬が 195 を 10 回連続で超えるまで学習を繰り返した．また，学習完了後に学習済みモデルで 100 エピソード分実験し，制御規則を作成するための入出力データを得た．

3　ファジィ制御

ファジィ制御は，入出力の関係をファジィ集合を用いた制御規則で表現し，制御規則の表現自体に曖昧さを含むことを許容することにより，数式で表現しにくい制御対象などにも頑健な制御を実現可能とする手法である [2]．ファジィ制御器は，与えられた入力を制御規則における前件部のメンバーシップ関数で評価され，その適用度を後件部に伝え，複数の制御規則の後件部の値の重みつき和を出力する．DQN で構築された制御器から得られた入出力関係から，ファジィ制御規則を導き出すことで，DQN の内部挙動を捉え，ファジィ制御規則を言語モデリングすることで，その挙動を言語で説明可能とする．

学習済み DQN の入出力データをデータセットとし，以下の過程を通じてファジィ制御器を構築する．

1. ファジィ制御器が使用する入力を決める
2. 入力のメンバーシップ関数を同定する
3. ファジィルールを作成する

3.1　ファジィ制御器が使用する入力情報の選定

CartPole への様々な入力において，制御則に取り入れるべき入力情報を安川 [3] の手法に基づき決定する．初めに，データセットをランダムに 2 分割し，それぞれグループ A・グループ B とする．次に，グループ A,B それぞれにおいて，入力の種類ごとに分類モデルを作成する．作成した分類モデルを用いて，入力の種類ごとに，式（1）で定義される CR 値 (Criterion of Regularity) を計算する．

$$CR = \frac{1}{2}\left(\frac{\sum_{i=1}^{k_A} |y_i^A - y_i^{AB}|}{k_A} + \frac{\sum_{i=1}^{k_B} |y_i^B - y_i^{BA}|}{k_B}\right) \tag{1}$$

ここで，式（1）中の変数は以下となる．

k_A, k_B	:	グループ A,B のデータ数
y_i^A, y_i^B	:	グループ A,B の出力のデータ
y_i^{AB}	:	グループ B のデータで作成したモデルにグループ A のデータを入れたときの出力
y_i^{BA}	:	グループ A のデータで作成したモデルにグループ B のデータを入れたときの出力

1 入力から始め，CR 値は下がらなくなるまで入力の個数を増やして CR 値を計算する．この CR 値が小さい入力の組み合わせをファジィ制御器の入力とする．

3.2　前件部メンバーシップ関数の同定

本研究では，制御則のファジィ集合表現に台形型のメンバーシップ関数を以下の流れで設定する．まず，入力情報をその出力である action ごとに分けてヒストグラムを作り，そのヒストグラムの凸部分のみを取り出したグラフを作成して，そのグラフを台形に近似することでメンバーシップ関数を同定する（図 1 参照）．

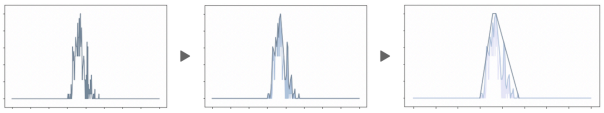

図 1: グラフの変化の様子

台形型のメンバーシップ関数は，データセットの 8 割のデータからなる制御則学習用データで作成した凸部分のみのグラフから台形の 4 つのパラメータの初期値を決め，パラメータを調整して作成する．台形のパラメータを左から順に p1,p2,p3,p4 とする．パラメータの調整は，はじめに「元の台形」「p1,p2,p3 を台形の幅の 1 %だけ右に動かした台形」「p2,p3,p4 を台形の幅の 1 %だけ左に動かした台形」の 3 つのメンバーシップ関数を作成する．次に，作成されたメンバーシップ関数を使用した 3 つのモデルそれぞれで定義される PI 値 (Performance Index) を算出する．データセットの個数を m，データセットにおける i 番目の入力に対する出力を y^i，i 番目の入力データに対するモデルの出力を $\hat{y}^i$ とすると，PI は式 (2) 以下のように定義される．

$$PI = \frac{1}{m}\sum_{i=1}^{m} |y^i - \hat{y}^i| \tag{2}$$

また，残り 2 割のテストデータを用いて，パラメータ調整前と調整後のモデルの精度を比較する．

3.3 ファジィルールの作成

後件部での出力値をシングルトンとしているため，制御に使用されるファジィルールはファジィ制御器が出力する action の種類と同じ数だけ作成される．ファジィルールの例を図 2 に示す．

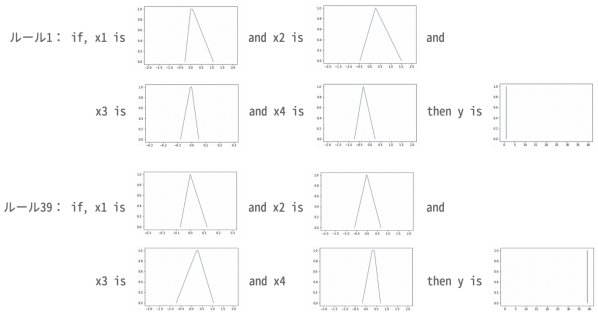

図 2: ファジィルールの例

3.4 ファジィ制御規則の言語モデリング

ファジィルールを言語モデリングすることにより言語化する．言語モデリングは小澤 [4] の方法を参考に行う．初めに「左に大」「右に小」「0 に近い」など入力の値に対する言語ラベルを設定する．次に 3.2 節で同定したメンバーシップ関数から各入力の言語ラベルを設定する．このとき，図 3 のように適用の度合いから出力する説明に「やや」「あたり」などの修飾語を加える．

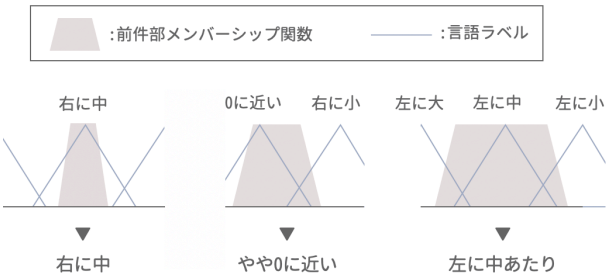

図 3: ファジィルールの例

4 実験

学習済みの深層強化学習モデルの内部挙動を言語化するために上述した手順でファジィ制御器を作成し、そのルールを出力する．併せて、そのルールで正しく CartPole を制御できるのかを検証する．

4.1 DQN での学習

今回は DQN で 342 エピソード学習したモデルの内部挙動を言語化する．このモデルの報酬関数のグラフを図 4 に示す．学習完了後に CartPole を 100 エピソード実験し、その入出力データを保存した．この 100 エピソードはどれも 200step 到達できていた．

4.2 ファジィ制御器の作成

4.1 にて得た入出力データを使用してファジィ制御器を作成する．このデータセットには「台車の位置」「台車の速度」「棒の角度」「棒の角速度」「action」のデータが 20,099 個入っている．

ファジィ制御器への入力の個数　入力の値から出力の値を分類するモデルを作成して CR 値を求めた．1 入力から 4 入力まで計算した結果, 4 入力の時が最も CR 値が小さくなったため，ファジィ制御器は 4 種類の入力を使用することにした．

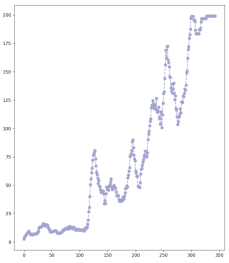

図 4: 平均報酬

前件部メンバーシップ関数
今回使用したデータセットでは 41 種類の action のうち 6 種類が実行されなかったため，35 種類の action に対して，1 つの action につき 4 つのメンバーシップ関数を同定した．テストデータを用いて PI 値を算出した結果，パラメータ調整前の PI 値が 10.7 であるのに対し，パラメータ調整後の PI 値が 6.3 であることからパラメータ調整が有効であることを確認した．

言語化されたファジィルール　以下のようなファジィルールが 35 個言語化された．以下の 2 つは図 2 で例として示したファジィルールを言語化したものである．

- 台車の位置が**右に小あたり**，台車の速度が**右に小あたり**，棒の角度が**やや左に小**，棒の角速度が**左に小あたり**のとき，action1 をとる
- 台車の位置が **0 に近いあたり**，台車の速度が **0 に近いあたり**，棒の角度が**左に小あたり**，棒の角速度が **0 に近いあたり**のとき，action39 をとる

4.3 生成された制御規則の検証

CartPole を 100 エピソード制御した結果，平均到達 step 数は 162.28step，200step 到達できたのは 38 エピソードであった．100 エピソード中 89 エピソードで 100step 到達していることから，DQN 学習後とまではいかないまでも CartPole を制御できていると言える．

4.4 考察

今回は台形型メンバーシップ関数のパラメータを逐次的に固定して調整を行なったことにより，パラメータ全体を考慮したチューニングがなされていないことが DQN での学習後の実験時と比較して制御精度が下がってしまったと考えられる．これに対して, 遺伝的アルゴリズム等を用いてパラメータの組み合わせ最適化を行うことで，制御の精度向上を実現できると考える．

5 おわりに

本研究では深層強化学習で学習されたモデルの制御規則を言語で説明することを試みた．今後の課題として，生成された制御規則を使用した制御の精度の上昇や，より制御が煩雑なタスクへの適用が挙げられる．

参考文献

[1] Mnih Volodymyr et al. Human-level control through deep reinforcement learning. *Nature*, Vol. 518, No. 7540, pp. 529–533, February 2015.

[2] 菅野道夫. ファジィ制御. 日刊工業新聞社, 1988.

[3] Takahiro Yasukawa. *A Fuzzy-Logic-Based Qualitative System Modeling*. PhD thesis, Tokyo Institute of Technology, 1994.

[4] 小澤順. ファジィ理論による数値情報の言語モデリング. Master's thesis, 東京工業大学, 1990.