F-044

ファジィ推論と方策勾配法とを融合した制御方式における逆強化学習の適用

Application of Inverse Reinforcement Learning in a Control System Integrating Fuzzy Reasoning and Policy Gradient Reinforcement Learning

小嶋 怜央† 石原 聖司†
Reo Kojima Seiji Ishihara

1. はじめに

ファジィ推論には推論過程の理解がしやすく，計算コストが少ないという利点がある．だが，ファジィ推論を行うにはルール重みやメンバシップ関数を事前知識に基づき定義する必要がある．これに対してファジィ推論と方策勾配法とを融合した制御方式の学習則[1]が提案されており，ルール重みやメンバシップ関数が方策勾配法による学習によって獲得可能であることが自動車の速度制御問題において実験的に示されている[2][3]．ただし，当該の制御モデルの学習を行うには報酬を定義する必要がある．これに対して，逆強化学習を当該の制御モデルに適用することにより報酬関数を自動獲得しながらメンバシップ関数の学習が可能なことが同様の問題において実験的に示されている[4]．ただし，文献[4]の実験では，使用するルールを先験的な知識に基づいて限定していた．また，逆強化学習を当該の制御モデルに適用した際のルール重みの学習は行われていなかった．

本研究では，使用するルールを限定せずに，文献[4]と同様の学習方法によって，ルール重みを固定した状態でのメンバシップ関数の学習結果と自動獲得された報酬関数およびメンバシップ関数を固定した状態でのルール重みの学習結果と自動獲得された報酬関数についてそれぞれ報告する．

2. 本研究での融合方式

エージェントの行動制御のファジィ制御ルールは次のように表すことができる．

Rule i：
if $(x_1\ is\ A_1^i)$ and ... and $(x_M\ is\ A_M^i)$
then $(y_1\ is\ B_1^i)$ and ... and $(y_N\ is\ B_N^i)$ with θ_i (1)

ここで，$i\ (=1,2,\cdots,n_R)$ はルール番号，$x_j\ (j=1,2,\cdots,M)$ は状態を記述する入力変数，$y_k\ (k=1,2,\cdots,N)$ は離散的な行動を記述する出力変数である．A_j^i / B_k^iはルールiのj / k番目の前件部/後件部に含まれる変数x_j / y_kを記述したファジィ表現で，それらのメンバシップ関数を$A_j^i(x_j) / B_k^i(y_k)$と表す．θ_i（≥ 0）はルールiの重みを表すパラメータである．

エージェントの各時刻の行動決定のために，次の目的関数

$$E(y;x,\theta,A,B) = -\sum_{i=1}^{n_R}\theta_i A^i(x)B^i(y) \quad (2)$$

の最小化を規範とした．ただし，入力値x/出力値yのルールiにおける前件部/後件部の適合度$A^i(x)\ /\ B^i(y)$を

$$A^i(x) \equiv \prod_{j=1}^{M} A_j^i(x_j) \quad (3)$$
$$B^i(y) \equiv \prod_{k=1}^{N} B_k^i(y_k) \quad (4)$$

で定義している．

時刻tにおける入力値$x(t)$に対する出力値$y(t)$の決定をエージェントの行動決定，すなわち，方策とし，式(2)の目的関数を持つ Boltzmann 分布による確率として次のように定義している．

$$\pi(y(t);x(t),\theta,A,B) \equiv \frac{e^{-\frac{E(y(t);x(t),\theta,A,B)}{T}}}{\sum_y e^{-\frac{E(y;x(t),\theta,A,B)}{T}}} \quad (5)$$

ただし，Tは温度と呼ばれるパラメータで，この方策で選択される出力値$y(t)$における目的関数$E(y(t))$の期待値をコントロールすることができる．すなわち，確率的な方策に基づく行動選択のランダム性の大きさを表している．

制御時には式(5)を重みとする重心y_G (t)を次のように定義して用いる．

$$y_G(t) \equiv \sum_y y\cdot\pi(y;x(t),\theta,A,B) = \langle y\rangle_{\pi(y)} \quad (6)$$

ただし，$\langle y\rangle_{\pi(y)}$は$\pi(y)$による$y$の期待値操作である．

方策勾配法により，エピソードごとの報酬期待値を極大化する，パラメータμに関する学習則は次のように与えられる．

$$\Delta\mu = \varepsilon\cdot\sum_{t=0}^{L-1} r(t)\cdot e_\mu(t) \quad (7)$$

ただし，Lはエピソード長，$r(t)$は報酬関数である．また，εは学習係数，e_μ(t)は特徴適性度

$$e_\mu(t) = \frac{\partial}{\partial\mu}\log\pi(y(t);x(t),\mu) \quad (8)$$
$$= -\frac{1}{T}\left[\frac{\partial E(y(t);x(t),\mu)}{\partial\mu} - \left\langle\frac{\partial E(y;x(t),\mu)}{\partial\mu}\right\rangle_{\pi(y)}\right] \quad (9)$$

である．

3. 逆強化学習

本研究では，報酬関数$r(t)$を次のように定義する．

$$r(t) = \boldsymbol{\tau}^T\cdot\boldsymbol{f}_{(x(t),y(t))} \quad (10)$$

ただし，$\boldsymbol{\tau} = (\tau_1,\cdots,\tau_u)^T$は報酬パラメータである．入力と出力の組を$\big(x(t),y(t)\big) \in F = \{1,\cdots,u\}$と$u$通りに量子化

† 東京電機大学 Tokyo Denki University

し，特徴ベクトル $\boldsymbol{f}_{(x(t),y(t))} = (0,\cdots,f_{(x(t),y(t))=h} = 1,\cdots,0)^T$ をu次元の one-hot ベクトルで定義する．本研究では，報酬パラメータ$\boldsymbol{\tau}$を最大エントロピー逆強化学習によって，

$$\Delta\boldsymbol{\tau} = \alpha\sum_{t=0}^{L}\left(\boldsymbol{f}^{*}_{(x(t),y(t))} - \boldsymbol{f}_{(x(t),y(t))}\right) \quad (11)$$

と更新する．ここで，αは学習係数，$\boldsymbol{f}^{*}$はエキスパートの特徴ベクトルである．つまり，$f_h < f_h^{*}$ではτ_hは増加し，$f_h > f_h^{*}$ではτ_hは減少する．これによって報酬関数の自動獲得を行う．

4. 自動車の速度制御問題への適用

一次元的な直線道路において，後方車両速度x_2の制御を行うことで，一定速度v_0で走行する前方車両との車間距離x_1を適切に保持する問題を考える．$t = 0$~80秒までに目標車間範囲$[l_0, l_1]$に入り，$t = 110$秒まで目標車間範囲内を保つことができればその問題は制御成功とする．ただし，制御を行う際のアクセルとブレーキの操作量yは，$y = -5.0 + 0.1h$（$h = 0,1,..100$）の 101 通りの離散値を使用する．また，学習時のエージェントの操作量yの決定には(5)式の方策$\pi(y)$を使用し，問題が制御成功か判定する際のエージェントの操作量は(6)式の重心y_Gを使用する．

4.1 メンバシップ関数

車間距離が“長い”/“短い”という 2 種類のファジィ表現のメンバシップ関数$A_1(x_1|l_0, l_1)$と後方車両速度が“速い”/“遅い”という 2 種類のファジィ表現のメンバシップ関数$A_2(x_2|v_0)$の 4 種類の前件部のメンバシップ関数をニューラルネットワークで表現する．アクセルを“強く”/“弱く”踏む，ブレーキを“強く”/“弱く”踏む，“なにもしない”の 5 種類の後件部のメンバシップ関数$B_1(y)$も学習可能だが，本研究では簡単のため，先験知識に基づいて定義したものを使用する．

4.2 量子化

本研究では，車間距離x_1を“遠い”，“やや遠い”，“やや近い”，“近い”の 4 通り，後方車両速度x_2を“速い”，“やや速い”，“やや遅い”，“遅い”の 4 通り，そしてアクセルとブレーキの操作量yを$-5.5 + g \le y < -4.5 + g$（$g = 0,1,..10$）のように 11 通りに量子化する．よって全体では$u = 4 \times 4 \times 11 = 176$通りに量子化する．

5. メンバシップ関数の学習に対する実験

学習に使用する 16 問の問題と 20 通りのファジィ制御ルールは文献[2]と同様のものを使用する．ルール重みθを固定した状態で前件部のメンバシップ関数A_1, A_2の結合荷重を(7)式のパラメータμとして強化学習を行う．ルール重みは先験知識に基づく固定値を使用し，既定の更新回数毎に(11)式より報酬関数の更新を行う．全16問で制御成功するまで前件部のメンバシップ関数A_1, A_2の結合荷重の強化学習と報酬関数の更新を繰り返し行う．

6. ルール重みの学習に対する実験

学習に使用する 16 問の問題と 20 通りのファジィ制御ルールは文献[2]と同様のものを使用する．前件部のメンバシップ関数A_1, A_2を固定した状態でルール重みを表すパラメータθを(7)式のパラメータμとして強化学習を行う．前件部のメンバシップ関数A_1, A_2は先験知識に基づいて定義したものを使用し，既定の更新回数毎に(11)式より報酬関数の更新を行う．全16問で制御成功するまでルール重みを表すパラメータθの強化学習と報酬関数の更新を繰り返し行う．

7. まとめ

ファジィ推論と方策勾配法とを融合した制御方式へ逆強化学習を適用することによって報酬関数の自動獲得機能を追加し，メンバシップ関数の学習とルール重みの学習の 2 種類の実験を行った．ルール重みを固定した状態での前件部のメンバシップ関数の学習結果と自動獲得された報酬関数，前件部のメンバシップ関数を固定した状態でのルール重みの学習結果と自動獲得された報酬関数についてはそれぞれ口頭発表時に報告する．

参考文献

[1] H. Igarashi and S. Ishihara:, “An Algorithm of Policy Gradient Reinforcement Learning with a Fuzzy Controller in Policies”, International Journal of Artificial Intelligence and Expert Systems, Vol. 4, pp. 17-26, 2013.

[2] 石原　聖司，五十嵐　治一，“ファジィ制御ルールにより表現された方策を持つ方策勾配法：自動車の速度制御問題への適用”，知能と情報，Vol. 32，No.4，pp. 801-810，2020.

[3] 市毛　竣，五十嵐　治一，石原　聖司，“ファジィ制御と強化学習の融合 メンバーシップ関数とルール重みの学習”，人工知能学会全国大会論文集，pp. 1D1GS203，2022.

[4] 小嶋　怜央，石原　聖司，“ファジィ推論と方策勾配法とを融合した制御方式における逆強化学習による報酬関数の自動獲得”，電子情報通信学会総合大会予稿集，pp.148，2024.