原著論文

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†

石原 聖司*¹・五十嵐 治一*²

ファジィ制御ルールによる行動決定のための方策の表現と方策勾配法による強化学習とを融合し、ファジィ制御ルールにおけるルール重みをエピソードあたりの報酬の期待値が極大となるように学習する方式が提案されている。この融合方式が自動車の速度制御の問題へ適用された研究例では、得られた結果の中に良好な解が見つかった一方、速度の望ましくない変動を生じる解も多く含まれていた。一般に、出力値に急な時間変化や振動をもたらす方策は望ましくなく、時間的に滑らかな出力値を与える方策が望ましい事例は多数考えられる。そこで本研究では、学習結果における出力値の急な変動を抑制するための改善策として、重心モデルによる非ファジィ化と時間に関する滑らか拘束項とを導入した目的関数を提案し、この融合方式に関してその場合の学習則を示す。さらに、報酬の与え方による出力値の変動に対する影響についても検討する。自動車の速度制御問題への適用実験の結果、提案した改善法について、出力値の望ましくない時間変動を抑制する効果があることを確認した。報酬関数については、設計の仕方によって学習の結果に悪影響が出る場合のあることを確認した。

キーワード：ファジィ制御、強化学習、方策勾配法、重心モデル、速度制御問題

1. はじめに

近年、深層学習の発展がめざましく、古典的な TV ゲーム [1] や囲碁 [2] などのゲームエージェントやロボットの行動制御にも適用されている [3]。しかし、大規模なニューラルネットワークモデルを精度良く学習させるには、大量の学習用データと計算資源が必要である [4]。しかも、学習により得られた特徴量は何であるのか、どういう制御ルールが獲得されたのかなど、学習で得られた制御の過程を人間が理解するには深層ネットワークの解析が必要となる。その点、ファジィ制御システムは、ファジィ制御ルール自体が人間の持つ知識表現として理解しやすく、システムの推論過程がわかりやすい。かつ、学習時や制御時の計算コストも低く抑えることができる。

このようなファジィ制御システムをベースにして、強化学習により最適な制御を実現しようとする研究が、以前から多く試みられてきた。両者の融合のメリットとしては、ファジィ制御の立場からは、if-then ルール内のメンバシップ関数の値や形状を記述するパラメータの値を教師データがない場合でも自動調整できるとい点が挙げられる。逆に、強化学習の立場からは、状態／行動変数の連続化や学習の階層化においてはファジィ集合による表現が適しているとの期待がある。

その中で、Igarashi と Ishihara は、ファジィ制御ルールによ

る行動決定のための方策の表現と、強化学習の中でも方策勾配法とを融合した学習方式を提案した [5]。この融合方式が自動車の速度制御の問題へ適用された研究例では、得られた結果の中に良好な解が見つかった一方、速度の望ましくない変動を生じる解も多く含まれていた。本論文では、この点についての改善法として、Igarashi と Ishihara の融合方式をベースに、重心モデルによる非ファジィ化と時間に関する滑らか拘束項とを導入した目的関数を提案し、その場合の学習則を示す。さらに、報酬の与え方の違いによる結果の差についても報告する。

本論文の構成は以下の通りである。2 節ではこれまでのファジィ制御と強化学習の融合方式について関連研究などをまとめた。3 節では本研究でベースとした融合方式について述べた。4 節では提案する重心モデルによる非ファジィ化と時間に関する滑らか拘束項の導入について述べた後、新たに導出したその学習則を記した。5 節では本研究で取り扱った速度制御問題の事例を説明した。6 節では計算機実験の内容と結果について詳述し、実験結果に対する考察も行った。7 節は本論文の全体のまとめである。また、4.3 節で述べた学習則の導出過程を本文の後に付録として記した。

2. 関連研究

2.1 ファジィ制御と強化学習の融合方式

ファジィ制御と強化学習の融合方式の研究は、用いた強化学習の種類により、大きく二つに分けることができる。マルコフ決定過程（Markov Decision Process：MDP）を仮定した価値ベースの強化学習法を用いる方法 [6–9] と、方策を直接学習する方策勾配法と呼ばれる方策ベースの強化学習法を用いる方法 [5, 10] である。

前者の制御システムは、Q 学習と組み合わせた場合が多い。その場合は、ルールの前件部には状態が記述され、後件部には

† Policy Gradient Reinforcement Learning for Policy Represented by Fuzzy Rules: Application to Simulations on Speed Control of an Automobile

Seiji ISHIHARA and Harukazu IGARASHI

*1 東京電機大学理工学部
School of Science and Engineering, Tokyo Denki University

*2 芝浦工業大学工学部
College of Engineering, Shibaura Institute of Technology

行動ごとの Q 値に対応するパラメータの値 [6,7] や関数 [8] が記述されるのが一般的である。しかし、前件部の状態を表す入力変数にはファジィ表現が用いられているのに対して、後件部の行動を表す出力変数にはファジィ表現が許されておらず、かつ離散的な行動だけを扱っていた。また、ルール自体の信頼性を記述した重みパラメータもなく、ルール自体の強化・生成や抑制・除去といった機能もなかった。

一方、後者の制御システムは、方策のパラメータを直接学習する方策勾配法と呼ばれる一連の強化学習法 [11-13] とファジィ制御とを組み合わせる。方策勾配法の一つである GPOMDP 法 [12] を用いた融合方式が Wang ら [10] により提案されているが、前件部のメンバシップ関数は固定され、後件部で記述された行動はファジィ表現が許されておらず、かつ、ルール重みも考慮されていない。これに対して、Igarashi と Ishihara が提案した融合方式 [5] では、ファジィ制御ルールの前／後件部におけるメンバシップ関数とルール重みを、エピソードあたりの報酬の期待値が極大となるように学習を行うことができる。従って、前／後件部のファジィ表現に関する自由度が高く、ルール自体の信頼性を記述した重みパラメータによるルールの強化や抑制の機能も備えているという特長がある。

2.2 制御問題への適用事例

Igarashi と Ishihara の提案した融合方式を、杉本らは RoboCup 小型リーグにおけるサッカーエージェントの行動決定の問題に適用した [14]。ここでは、システム出力と人間の行動決定とを一致させるという教師あり学習問題へ適用を試み、学習アルゴリズムの有効性を示した。ただし、この事例では、行動（出力変数）がシュート・パス・ドリブルの三つに限定されており（離散変数）、最終的にはこの三つの行動から一つを選ぶという判断を行うだけであった。¹すなわち、サッカー局面の静止画像を 1 シーンずつ入力し、その場における人間の判断と確率的に近い行動決定を行うための方策を学習させた。従って、実際にシュートなどの行動を行って環境の状態を変化させ、その結果に対して与えられた報酬を行動決定へフィードバックする強化学習の事例ではなかった。

強化学習をとまなう制御問題へ適用した事例としては、土屋らの研究がある [15]。これは、一定速度で走行する先行車に対して、一定の範囲の車間距離を維持させるように後続車の速度を制御する問題において、ファジィ制御ルールの重要性を表すパラメータ（ルール重み）を学習した事例である。この制御システムの出力変数は、アクセルやブレーキの強さを表す連続変数である。この研究では、得られた重みの解は、人間が手作業で選んだ重みの値や、ファジィ制御における代表的手法である Min-Max 重心法²による最良解と比べて同等以上の制御精度を与えた。しかし、速度に振動を生じる不適切な解³を与えることも多く、学習の目的関数の正則化や、報酬の与え方について

の検討が今後必要であると結論付けられていた。

一般に、エージェントの行動に対応する出力が連続変数の値で表される場合、出力値に急な時間変化や振動をもたらす方策は望ましくなく、時間的に滑らかな出力値を与える方策が望ましい事例は多数考えられる。そこで、本論文では、行動に対応する出力が連続変数である事例に Igarashi と Ishihara の融合方式を適用する際に、目的関数の正則化（重心モデルと滑らか拘束項の導入）と報酬関数の設計見直しとによる改善で、出力値の急激な変化や振動がある望ましくない解の発生を抑えるようにすることを考えた。

3. 本研究のベースとなった融合方式

本研究の制御システムは Igarashi と Ishihara の融合方式 [5] をベースとしている。この融合方式で示されたファジィ制御ルールを 3.1 節、目的関数と方策を 3.2 節、ルール重みの学習則を 3.3 節で説明する。なお、この融合方式で示されたメンバシップ関数の学習則については本研究では取り扱わない。

3.1 ファジィ制御ルール

エージェントの行動の制御に使用されるファジィ制御ルールは、

Rule i :

$$\begin{aligned} &\text{if } (x_1 \text{ is } A_1^i) \text{ and } \dots \text{ and } (x_M \text{ is } A_M^i) \\ &\text{then } (y_1 \text{ is } B_1^i) \text{ and } \dots \text{ and } (y_N \text{ is } B_N^i) \text{ with } \theta_i \end{aligned} \quad (1)$$

と表される。ここで、 $i = 1, 2, \dots, n_R$ はルール番号、 x_j ($j = 1, 2, \dots, M$) は状態を記述する入力変数、 y_k ($k = 1, 2, \dots, N$) は行動を記述する出力変数である。 A_j^i / B_k^i はルール番号 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が $A_j^i(x_j) / B_k^i(y_k)$ である。 θ_i (≥ 0) はルール重みを表すパラメータである。本研究ではこの θ_i を学習の対象とする。

3.2 目的関数と方策

エージェントの各時刻における行動決定のための規範となるのは、目的関数

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

の最小化である。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。式 (3), (4) は、それぞれ、前件部／後件部が論理積 (AND) からなる推論ルールにおいて、入力値 x / 出力値 y が前件部／後件部の命題を支持している度合いを、各次元のファジィ集合の適合度の算術積で表すという考え方である。

1 各行動を選択する度合いが後件部に記述されている。

2 ルールの種類は人間が手作業で選択する。

3 細かい速度振動は人間に不快感を与え、実際の制御も容易ではないと思われる。文献 [15] では図 5 にその例が示されている。本論文でも具体例を 6.5 節の図 6 で示した。

従って、式 (2) の目的関数は入力値が x のときに、各ルールにより出力値 y が適切かどうか支持される度合いの統合結果を表している。

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（学習システム）の行動決定、すなわち、方策とみなす。この方策は、式 (2) の目的関数を持つ Boltzmann 分布

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

による確率として定義される。ただし、 T は温度と呼ばれるパラメータであり、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す。

3.3 方策 π によるルール重み θ_i の学習則

ルール重み θ_i に対する学習則は、

$$\Delta\theta_i = \varepsilon r \sum_{t=0}^{L-1} e_{\theta_i}(t) \quad (6)$$

と表される。ここで、 L はエピソード長、 r は報酬関数、 ε (> 0) は学習係数をそれぞれ表す。また、 $e_{\theta_i}(t)$ は特徴的適正度 (characteristic eligibility) と呼ばれ、

$$\begin{aligned} e_{\theta_i}(t) &\equiv \frac{\partial}{\partial \theta_i} \ln \pi(y(t)) \\ &= \frac{1}{T} A^i(x(t)) \left\{ B^i(y(t)) - \langle B^i(y) \rangle_{\pi(y)} \right\} \end{aligned} \quad (7)$$

で与えられる。ただし、

$$\langle X(y) \rangle_{\pi(y)} \equiv \sum_y X(y) \pi(y) \quad (8)$$

である。

杉本らは、人間の考えた望ましい出力分布に確率の方策 π が近づくように、ルール番号 i の $A_j^i(x_j)$ ($j = 1, 2, 3$) に含まれるパラメータ 6 種類、 $B_k^i(y_k)$ ($k = 1, 2, 3$) の値 3 種類、および、ルール重み θ_i のそれぞれを Igarashi と Ishihara の融合方式で示された学習則によって学習した [14]。本研究とは異なり、以降の節で示す重心モデルや滑らか拘束項は扱っていない。

4. 推論方式と学習法の改善

4.1 重心モデルを用いた非ファジィ化

土屋らは、確率の方策を利用して変数 y の離散値から重心を代表値として取ることににより連続的な値を出力する非ファジィ化のモデルを考えた [15]。従来、Min-Max 演算を代数積-加算演算に置き換えたファジィ推論のモデルとして、「代数積-加算-重心モデル」[16, 17] が提案されている。両モデルは類似しているが、後件部のファジィ集合の代数和 $B(y) = \sum_k B_k(y)$ の幾何学的な重心位置

$$y_g \equiv \frac{\int y B(y) dy}{\int B(y) dy} \quad (9)$$

を後者のモデルでは求める一方、式 (2) のルール重み $\{\theta_i\}$ が

反映された式 (5) を重みとする期待値操作に従って代表値となる重心 $y_G(t)$ を前者のモデルでは得るという点が異なる。すなわち、

$$\begin{aligned} y_G(t) &\equiv \sum_y y \pi(y; x(t), \theta, A, B) \\ &= \langle y \rangle_{\pi(y)} \end{aligned} \quad (10)$$

と定義する。

土屋らは、 $\{\theta_i\}$ を学習するフェーズでは、時刻 t の行動に対応する出力値を式 (5) の確率分布に従って変数 y の離散値から確率的に選択した。一方、学習によって獲得した $\{\theta_i\}$ を評価するフェーズでは、時刻 t の行動に対応する出力値を決定論的に式 (10) の重心値とした。本研究とは異なり、以降の節で示す、重心モデルを学習のフェーズへ導入するための目的関数や学習則、および、滑らか拘束項は扱っていない。

4.2 重心モデルと滑らか拘束項を導入した学習フェーズ

本研究では、式 (10) による時刻 t の重心 $y_G(t)$ と変数 y の離散値との 2 乗誤差を、出力の時間変化に関する滑らかさの拘束条件の下で最小化する次の目的関数を提案する。

$$E'(y) = \frac{1}{2} (y - y_G(t))^2 + \lambda (y - y(t-1))^2 \quad (11)$$

式 (11) 右辺第 1 項により y が重心に近いほど、また、第 2 項により y が 1 時刻前の値に近いほど、目的関数 $E'(y)$ の値は小さくなる。すなわち、重心の値からと直前の値からの乖離を小さくすることで、 y の時間変動を抑える機能を持つ。ただし、 λ (≥ 0) は第 2 項による拘束条件の強さを表すパラメータである。

ここで、式 (11) の目的関数の最小化を規範とする確率の方策

$$\pi'(y(t)) \equiv \frac{e^{-\frac{E'(y(t))}{T'}}}{\sum_y e^{-\frac{E'(y)}{T'}}} \quad (12)$$

を定義する。

本研究では、学習法の改善として、 $\{\theta_i\}$ を学習するフェーズに式 (12) の確率の方策を導入する。すなわち、時刻 t の行動に対応する出力値を式 (12) の確率分布に従って変数 y の離散値から確率的に選択する。さらに、学習によって獲得した $\{\theta_i\}$ を評価するフェーズに式 (11) の最適解

$$y_0(t) \equiv \frac{y_G(t) + 2\lambda y(t-1)}{1 + 2\lambda} \quad (13)$$

を導入する。すなわち、時刻 t の行動に対応する出力値を決定論的に式 (13) の $y_0(t)$ とする。上述のように、本論文で提案する改善法は、出力の急な変動を抑制する機能を持つ式 (11) の目的関数を、学習フェーズと評価フェーズにそれぞれ適用したものである。この改善法の学習フェーズにおける学習則は次節で示す。

4.3 方策 π' によるルール重み θ_i の学習則

ルール i の重み θ_i についての方策 π' による学習則は、方策勾配法の学習則 [11, 13] から、

$$\Delta\theta_i = \varepsilon' r \sum_{t=0}^{L-1} e'_{\theta_i}(t) \quad (14)$$

と表される。ここで、 $\varepsilon' (> 0)$ は学習係数を表す。特徴的適正度 $e'_{\theta_i}(t)$ は、式 (11) の目的関数と式 (12) の方策から、

$$\begin{aligned} e'_{\theta_i}(t) &\equiv \frac{\partial}{\partial \theta_i} \ln \pi'(y(t)) \\ &= -\frac{A^i(x(t))}{TT'} \langle (y - y_G(t)) B^i(y) \rangle_{\pi(y)} \left(y(t) - \langle y \rangle_{\pi'(y)} \right) \end{aligned} \quad (15)$$

与えられる。式 (14) と (15) より、前／後件部の適合度や報酬が大きいルールの重みは更新量が大きくなることと、重心値や直前時刻での値の近傍から実際に選ばれた出力値 $y(t)$ と確率的方策 π' によるその期待値 $\langle y \rangle_{\pi'(y)}$ との差に応じて重みが更新されることがわかる。式 (15) の導出過程は付録に記した。なお、重みの更新はエピソード終了ごとに行う。

5. 制御問題の事例

本節では、土屋らの研究 [15] でも取り上げられた、本研究で取り扱う自動車の速度制御問題の事例について述べる。

5.1 自動車の速度制御問題

2 台の自動車が一次元的な直線道路を同じ方向に走行しているとす。ただし、前方車両速度は一定であるとする。2 台の車間距離（前方車両に対する後方車両の追従距離）が、目標とする制御範囲内に所定の時間内で収まり、さらに、その範囲内に一定時間以上とどまるように、アクセルとブレーキを用いて後方車両の速度を制御する問題を考える。

5.2 ファジィ制御ルール

後方車両の速度を制御するために以下のファジィ制御ルールを用いた。

$$\begin{aligned} \text{Rule } i : & \text{if } (x_1 \text{ is } A_1^i) \text{ and } (x_2 \text{ is } A_2^i) \\ & \text{then } (y_1 \text{ is } B_1^i) \text{ with } \theta_i \end{aligned} \quad (16)$$

式 (16) のルールの前件部の入力変数 $x = (x_1, x_2)$ には（車間距離，後方車両速度）を対応させ、後件部の出力変数 y_1 には後方車両のアクセルおよびブレーキの操作量を対応させる。また、車間距離 x_1 を“長い／短い”，後方車両速度 x_2 を“速い／遅い”の各 2 通りに分類する。さらに、アクセルとブレーキの操作量 y_1 を“アクセルを強く／弱く踏む”，“ブレーキを強く／弱く踏む”，“何もしない”の 5 通りに分類する。これらの組み合わせで得られるすべてのルール（ $2 \times 2 \times 5 = 20$ 個）を用意した（表 1 参照）。

これら 20 個のルールは機械的に生成しているために、直感的に不適切なルールも含まれている。例えば、ルール 8 は前方車両との車間距離が長く、自車（＝後方車両）の速度が遅いの

表 1 速度制御問題で使したファジィ制御ルール（ac. は加速（acceleration），de. は減速（deceleration））

Rule	Following distance	Speed of the following car	Operation of the following car
1	long	fast	strong ac.
2			weak ac.
3			strong de.
4			weak de.
5			none
6		slow	strong ac.
7			weak ac.
8			strong de.
9			weak de.
10			none
11	short	fast	strong ac.
12			weak ac.
13			strong de.
14			weak de.
15			none
16		slow	strong ac.
17			weak ac.
18			strong de.
19			weak de.
20			none

目的は、これらの不適切なルールの重みを抑制する一方、適切なルールの重みを強化して、所望の制御を達成する方策を得ることである。

5.3 メンバシップ関数の設定

式 (16) のルール中のメンバシップ関数（車間距離に関する $A_1^i(x_1)$ ，後方車両速度に関する $A_2^i(x_2)$ ，操作量に関する $B_1^i(y_1)$ ）を図 1～3 のように設定した。図 1 に示すように、車間距離が“長い／短い”というファジィ表現を目標範囲 $[l_1, l_2]$ を用いて折れ線で表現した。図 2 では、後方車両速度が“速い／遅い”というファジィ表現を、前方車両速度 v を用いて定義した。図 3 では、アクセルとブレーキを“強く／弱く踏む”というファジィ表現のメンバシップ関数を定義した。ただし、“何もしない”という行動出力値は $y_1 = 0$ のときに 1，そうでないときに 0 という非ファジィ表現を用いた。なお、図 3 のメンバシップ関数は、アクセルとブレーキとは機構が異なるという想定の下、操作量 y_1 に対応するファジィ表現の特性がアクセルとブレーキとで異なる場合の例として設定したものである。そのため、 $y_1 = 0$ を中心に左右非対称な形状になっている。

時間については離散時間（ $\Delta t = 1 \text{ sec}$ ）を考える。なお、 $\Delta t = 1 \text{ sec}$ は、人間の操作間隔を想定した粗い制御間隔の一例であるが、時間間隔の長い粗い制御が制御目標の達成を容易にして問題を単純化するわけではない。アクセルを最大に踏み続けたときに（ $y_1 = 5$ ），10 秒で 100 km/h の速度に到達するように、アクセルの操作量 y_1 のときに加速度が $2y_1 \text{ [km/h/s]}$ だけ車両に加わると仮定した。ブレーキ操作時も同様に減速のために $-2y_1 \text{ [km/h/s]}$ の加速度が加わると仮定した。すなわち、操作量 y_1 と加速度との対応関係については、図 3 のメン

図 1 車間距離 x_1 に関するメンバシップ関数 $A_1^i(x_1)$

図 2 後方車両速度 x_2 に関するメンバシップ関数 $A_2^i(x_2)$

図 3 操作量 y_1 に関するメンバシップ関数 $B_1^i(y_1)$

バーシップ関数とは独立に定義できることから、加速時と減速時とで加速度の絶対値が同一となる単純化した線形な例を使用した。

6. 計算機実験

表 2 に示す 3 通りの適用法それぞれについて、5 節で説明した自動車の速度制御問題に対する学習及び評価の実験を行った。(i) は改善前の従来法であり、土屋らの方法 [15] と同様である。(ii) と (iii) が本論文で提案する改善法である。ただし、(ii) は改善法における滑らか拘束項を使用しない方法である。いずれの適用法においても、変数 y_1 の値には離散値 ($y_1 = -5.0 + 0.1h$ ($h = 0, 1, \dots, 100$)) を使用した。なお、(iii) の適用法においては、 λ の値の大まかな調整を行った結果、経験的に $[0.05, 0.1]$ の値が良いことがわかった。次に、0.05 から 0.01 刻みに予備的な学習実験を行ったところ、学習結果に大きな差が見られなかった。そこで、今回は、 $\lambda = 0.06$ と設定し、本実験を行った。

6.1 学習実験の方法

車両速度および車間距離の初期値、車間距離の目標範囲

表 2 3 通りの適用法の目的関数・出力・学習則・拘束条件

Meth.	Learning phase			Evaluation phase	Constraint condition
	Objective function	Output	Learning rule	Output	
(i)	$E(y)$	$\pi(y)$	(6) & (7)	y_G	—
(ii)	$E'(y)$	$\pi'(y)$	(14) & (15)	y_O	$\lambda = 0$
(iii)					$\lambda = 0.06$

表 3 各問題の初期条件の値

Problem	Conditions				
	Initial speed (km/h)		Initial following distance (m)	Target distance (m)	
	Forward car	Following car		l_1	l_2
1	20	30	50	30	45
2	20	30	50	10	15
3	20	30	10	30	45
4	20	30	10	10	15
5	30	30	50	30	45
6	30	30	50	10	15
7	30	30	10	30	45
8	30	30	10	10	15
9	50	30	50	30	45
10	50	30	50	10	15
11	50	30	10	30	45
12	50	30	10	10	15
13	60	30	50	30	45
14	60	30	50	10	15
15	60	30	10	30	45
16	60	30	10	10	15

$[l_1, l_2]$ の値を表 3 に示すように設定した 16 通りの問題を学習用に用意した。これらの問題に対して、表 1 のルール 1~20 に対応する重み θ_i ($i = 1, \dots, 20$) を、方策勾配法により強化学習する実験を (i) ~ (iii) の 3 通りの適用法それぞれで行った。

後方車両の速度制御は、離散時間である $t = 0$ から開始して $t = 110$ で終了した。すなわち、1 エピソードの長さ $L = 110$ である。ただし、車間距離が 0 m 未満となった（追突した）場合と、車間距離が 200 m 以上離れた場合については、その時点で制御を打ち切りエピソード終了とみなした。車間距離と時間に関する次の二つの要件、

要件 1. 車間距離が目標範囲 $[l_1, l_2]$ に収まる。

要件 2. $80 \leq t \leq 110$ の区間で要件 1 が常に成立する。

を満たすことを速度制御の目標とした。⁴

報酬関数については、土屋らの研究で取り扱われたタイプ [15] と同様である賞罰を使い分けるタイプ ($r = r_2$) に加えて、罰のみを使うタイプ ($r = r_1$) の 2 通りを用意した。すなわち、(i) ~ (iii) の各適用法において、 $r = r_1$ の場合と $r = r_2$ の場合との 2 通りの学習結果を得た。本研究では、制御結果の各ケースにおいて、 r_1 および r_2 の値を次のように設定した。

4 これら二つの要件を最短時間で実現することまでは要求されていない。

ケース 1. $t = t_{in}$ (≤ 80) に目標範囲に入り、その状態を $t = 110$ まで維持した場合（目標達成）：

$$r_1 = 0, \\ r_2 = 0.01/(t_{in} + 1).$$

ケース 2. $t = t_{in}$ (> 80) に目標範囲に入り、その状態を $t = 110$ まで維持した場合（時間の要件未達成）：

$$r_1 = (80 - t_{in})/100 + c, \\ r_2 = 0.01/(t_{in} + 1).$$

ケース 3. $t = 110$ となった段階で目標範囲に入っていない場合（車間距離・時間の要件ともに未達成）：

$$r_1 = r_2 = -|((l_1 + l_2)/2 - x_1)/20000| + c.$$

ケース 4. 車間距離が 0 m 未満となり、その時点で制御を打ち切った場合（車間距離・時間の要件ともに未達成）：

$$r_1 = r_2 = -x_1^2/100 + c.$$

ケース 5. 車間距離が 200 m 以上となり、時刻 $t = t_{far}$ で制御を打ち切った場合（車間距離・時間の要件ともに未達成）：

$$r_1 = r_2 = (t_{far} - 110)/100 + c.$$

ここで、 x_1 は各ケースにおける制御終了時の車間距離であり、 $c = -0.01$ とした。目標を達成した場合、 r_1 では報酬を与えない一方、 r_2 では目標範囲に入った時間が早いほど大きなプラス報酬を与えるという点で両者は異なる。なお、車間距離・時間の要件ともに目標を達成しなかった場合、車間距離や時間に依じてマイナスの報酬を与えるという点は両者に共通である。すなわち、 r_1 は、失敗する方策をとらないようにすることを想定した報酬関数である。一方、 r_2 は、失敗する方策をとらないだけでなく目標範囲に入る時間が次第に早くなるようにすることも含めて当該の方策を強化することを想定した報酬関数である。

いずれの適用法と報酬関数においても、初期条件を設定した後の学習手順は、

手順 1. 学習フェーズの行動決定により学習用問題に対するエピソードを一つ求め、学習則に従ってルール重みを更新する。ただし、 $\sum_i \theta_i = 1.0$ ($\theta_i \geq 0$) となるようにルール重みは正規化する。これを、問題 1 から順に逐次行い、問題 16 を終えたら手順 2 に進む。

手順 2. 評価フェーズの行動決定により学習用問題に対するエピソードを一つ求め、制御目標の達成を確認する。これを、問題 1 から順に逐次行い、問題 16 を終えたら手順 3 に進む。

手順 3. 学習用問題のすべてで制御目標を達成できたら、そのルール重みを結果に採用して学習を終了する。制御目標を達成できない問題があれば、手順 1 に戻る。ただし、この手順の反復回数（以降、学習回数と呼ぶ）が 200 回に到達した場合には、その時点でのルール重みを結果に採用して学習を打ち切って終了する。

とした。なお、初期条件については、すべての適用法において θ_i の初期値を 0.05、(i) の適用法において $T = 0.04$ 、 $\varepsilon = 0.0075$ 、(ii) と (iii) の適用法において $T = T' = 0.04$ 、 $\varepsilon' = 0.0003$ とそれぞれ設定した。初期条件の設定からこの学習手順の終了ま

表 4 学習実験 10,000 回の結果

	$r = r_1$				$r = r_2$			
	$ S $	$ S_c $	$\overline{m_c}$	$\overline{t_{in_c}}$	$ S $	$ S_c $	$\overline{m_c}$	$\overline{t_{in_c}}$
(i)	9873	9036	25.1	17.6	9890	8673	19.3	17.4
(ii)	9921	9087	36.8	16.8	9975	9040	16.7	16.5
(iii)	9908	9539	36.4	17.0	9951	9686	18.8	17.2

で実験回数 1 とカウントする。すなわち、実験 1 回ごとに、学習済みのルール重み θ_i ($i = 1, \dots, 20$) を 1 セット得る。本研究では、10,000 回の実験を行い、学習済みのルール重みを 10,000 セット得た。また、学習前のルール重み ($\theta_i = 0.05$) を用いて、評価フェーズの行動決定による速度制御を行った場合、学習用のいずれの問題に対しても制御目標を達成しないことをすべての適用法で確認済みである。

6.2 学習実験の結果と考察

学習用のすべての問題に対して制御目標を達成した学習済みルール重みのセット $s = (\theta_1, \dots, \theta_{20})$ を要素とする集合を S と定義する。後方車両の追従速度の急激な変動が抑制されていることを $t = 110$ の時点で判定するための簡易的な条件、

条件 1. $t = 109$ と $t = 110$ での後方車両速度の差（加速度）の絶対値が 0.1 km/h 未満である。

条件 2. $t = 110$ のとき、前方車両の速度と後方車両の速度との差の絶対値が 0.1 km/h 未満である。

を設定し、条件 1 と条件 2 を共に満たすエピソードを生成した s を要素とする集合を $S_c \subseteq S$ と定義する。前方車両の速度が一定である場合、走行の快適性や効率性の観点から後方車両の追従速度の変動は少ない方が一般的に望ましい。そこで、各 $s \in S_c$ の獲得に要した学習回数の平均 $\overline{m_c}$ 、各 $s \in S_c$ を用いた制御によって後方車両が各問題の目標範囲に入るまでの時間ステップ数の平均 $\overline{t_{in_c}}$ をそれぞれ実験データから求めた。学習実験の結果としてこれらの値を表 4 にまとめて示した。

表 4 の $|S|$ より、(i) ～(iii) のいずれの適用法においても、制御目標を達成する解（ルール重みのセット）を 98% 以上の実験で学習できたことがわかる。一方、表 4 の $|S_c|$ より、追従速度の変動を抑制する良質な解については、 r_1 と r_2 の報酬関数によらず、(i) に対して (ii)、(iii) の順に増加したことがわかる（表 4 灰色のセル参照）。この結果から、4.2 節で述べた改善法について、出力の時間変動を抑制する機能の効果が学習実験において確認された。

次に、二つの報酬関数について比較する。表 4 の $\overline{m_c}$ より、 r_1 に対して r_2 は学習回数を減少させる効果があったことがわかる。一方、表 4 の $\overline{t_{in_c}}$ より、車間距離が目標範囲に入るまでの時間ステップ数の r_2 による短縮効果は大きいとは言えない。また、表 4 の $|S_c|$ より、 r_1 に対して r_2 では追従速度の変動を抑制する良質な解が減少してしまう場合もあることがわかる。

6.3 評価実験の方法

学習用の問題とは初期値を変更し、前方／後方車両の速度 (km/h) を $\{45, 55, 65, 75, 85\} / \{0, 10, 30, 50, 70\}$ 、車間距

離 (m) を $\{30, 45, 65, 70, 80\}$, 車間距離の目標範囲 (m) を $\{[10, 30], [20, 40], [40, 60], [50, 60], [60, 70]\}$ からそれぞれ設定してそれらを組み合わせて $5 \times 5 \times 5 \times 5 = 625$ 通りの問題を用意した. さらに, 前方/後方車両の速度 (km/h) を $\{40, 50, 60\} / \{20, 40, 60\}$, 車間距離 (m) を $\{20, 40, 60\}$, 車間距離の目標範囲 (m) を $\{[10, 20], [40, 50], [45, 60]\}$ からそれぞれ設定してそれらを組み合わせて (前方車両速度 50 km/h と車間距離 20 m との組み合わせを除く) $3 \times 3 \times 3 \times 3 - 3 \times 3 = 72$ 通りの問題も用意した. これら合計 697 通りのテスト用 (未学習) 問題に対して, 学習実験で獲得したルール重みの各セット $s \in S_c$ を用いた行動決定の評価を行った. すなわち, 評価フェーズの行動決定によりテスト用問題に対するエピソードを求め, 制御目標の達成を調べた. また, 学習前のルール重み ($\theta_i = 0.05$) を用いて, 評価フェーズの行動決定による速度制御を行った場合, テスト用のいずれの問題に対しても制御目標を達成しないことをすべての適用法で確認済みである.

6.4 評価実験の結果と考察

$s \in S_c$ のうち, テスト用のすべての問題に対して制御目標を達成した s を要素とする集合を $S' \subseteq S_c$ と定義する. さらに, $s \in S'$ のうち, 6.2 節の条件 1 と条件 2 を共に満たすエピソードを生成した s を要素とする集合を $S'_c \subseteq S'$ と定義する. 各 $s \in S'_c$ を用いた制御によって後方車両が各問題の目標範囲に入るまでの時間ステップ数の平均 $\overline{t'_{inc}}$ を実験データから求めた. 評価実験の結果としてこれらの値を表 5 にまとめて示した.

表 5 の $|S'_c|$ より, 追従速度の変動を抑制する良質な解については, 報酬関数によらず, 学習実験の結果と同様に (i) < (ii) < (iii) の順に多くなったことがわかる. さらに, 表 5 の $|S'_c|/|S_c|$ より, 学習用問題に対する良質な解がテスト用問題に対しても良質な解であった割合, すなわち, 汎用性のある良質な解が得られた割合についても同様の順に大きくなった. これらの結果から, 出力の時間変動を抑制する機能について, 4.2 節で述べた式 (11) の目的関数の有効性が未学習問題に対しても検証できた.

次に, 二つの報酬関数について比較する. 表 5 の $\overline{t'_{inc}}$ の値から, r_1 に対して r_2 は, 目標範囲に入るまでの時間ステップ数の短縮効果がそれほど大きくはないのに, 追従速度の変動を抑制する良質な解が減少してしまうという副作用が見られた. すなわち, 目標範囲に入るまでの時間短縮効果を意図して r_2 のようなタイプの報酬関数を設定すると, r_1 のようなタイプの報酬関数とは学習の目標自体が変わり, 悪影響が出る場合もあるため, 報酬関数の設計には注意が必要であると言える.

6.5 学習したルール重みと速度制御の例

$s \in S'_c$ のうち, 目標範囲に入るまでの時間ステップ数の全問題に関する平均が最小となる s の値を図 4 および図 5 に示した.

図 4 および図 5 より, 表 1 に示したルール 6 とルール 13 に対する重みが相対的に大きくなり, ルール 8 とルール 11 に対する重みが相対的に小さくなる傾向があったことがわかる. ルール 6 は, 車間距離が長く後方車両速度が遅いとき強く加

表 5 学習したルール重み $s \in S_c$ に対する評価実験の結果

	$r = r_1$				$r = r_2$			
	$ S' $	$ S'_c $	$ S'_c / S_c $	$\overline{t'_{inc}}$	$ S' $	$ S'_c $	$ S'_c / S_c $	$\overline{t'_{inc}}$
(i)	569	563	0.062	19.5	545	543	0.063	19.3
(ii)	670	668	0.074	20.0	576	573	0.063	19.6
(iii)	761	760	0.080	19.9	649	649	0.067	19.5

図 4 報酬関数 r_1 を使用して学習したルール重みの例

図 5 報酬関数 r_2 を使用して学習したルール重みの例

速するというルールであって目標達成のためには適切である. ルール 13 についても, 車間距離が短く後方車両速度が速いとき強く減速するというルールであるため同様に適切である. 一方, ルール 8 およびルール 11 は, 上述のルールとは反対の意味を持つものであり, 目標達成のためには不適切であると言える. すなわち, ルール 6・13 の重みが大きくなり, ルール 8・11 の重みが小さくなる組み合わせは汎用性の高い有効なものであると考えられる.

ルール 1~5 およびルール 16~20 については, その適切性が明確ではなく, 目標達成にとって必ずしも重要な役割を持つものではない. 図 4 および図 5 より, (ii) と (iii) では, これらの重要でない重みは初期値 (0.05) からあまり変化していないことがわかる. 一方, (i) では, 重要でない重みについても初期値からの増減が目立った. この違いは, 正則化機能の有無の差によるものであると考えられる. すなわち, 重心と滑らか拘束条件とを用いた式 (11) の目的関数の効果により, ルール重みの不要なばつきが抑制されたと考えられる.

s を用いた速度制御結果の例として, 表 3 の問題 1 に対する評価フェーズでの行動決定によって追従速度の変動が収まらな

図6 r_2 と (i) で学習した $s \notin S_c$ による速度制御の例

図7 r_2 と (i) で学習した $s \in S_c$ による速度制御の例

い場合 ($s \notin S_c$ の場合) と収まる場合 ($s \in S_c$ の場合) の追従速度・車間距離の変化の様子をそれぞれ図6と図7に示した。これらの図はいずれも、報酬関数 r_2 と (i) で学習した s による結果である。図6と図7より、車間距離が目標範囲(図中の点線 l_1 と l_2 とで挟まれた範囲)に入り制御目標を達成する解であっても、追従速度の変動が収まらない図6のような結果は、図7のように変動が収まる結果と比べると、快適性や効率性の点で適切とは言えないことがわかる。

7. まとめ

本論文では、出力となる行動の値が連続変数であり、かつ、行動決定に基づいてエージェントを実際に行動させ、その環境への影響に対して報酬を与えるファジィ制御問題の強化学習に対して、Igarashi と Ishihara の融合方式 [5] を適用することを考えた。ファジィ推論の方式には、行動に対応する出力変数の離散値から、確率的方策を用いた期待値操作で連続的な値である重心を求める非ファジィ化のモデル(重心モデル)を採用した。学習結果における出力値の急な変動を抑制するための改善策として、行動に対応する出力変数の離散値と重心との誤差を、出力値の時間変化に関する滑らかさの拘束条件の下で最小化する目的関数を提案し、その学習則を示した。

提案した改善法を、一定速度で走行する先行車に対して一定の範囲の車間距離を維持させるように後続車の追従速度を一定の時間内で制御する事例に適用し、ファジィ制御ルールの重要性(ルール重み)を表すパラメータの学習を行った。計算機実験の結果から、出力値の望ましくない時間変動を抑制する提案した改善法の効果を確認することができた。報酬関数について

は、目標を達成しない方策を取らないようにすることに加えて、目標範囲に入るまでの時間短縮効果を含めて設計をした場合、学習回数の減少効果がある一方で、出力値の急激な時間変動を持つ解を誘発しやすい傾向があることがわかった。報酬関数を設計する際には、このような副作用に注意して今後も検討が必要である。

本論文では取り扱わなかったが、ファジィ制御ルール中に含まれるメンバシップ関数の学習則について、今後検証していく予定である。

参考文献

- [1] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis: "Human-Level Control through Deep Reinforcement Learning," *Nature*, Vol.518, pp. 529-533, 2015.
- [2] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis: "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature*, Vol.529, pp. 484-489, 2016.
- [3] Y. Duan, X. Chen, R. Houthoofd, J. Schulman, and P. Abbeel: "Benchmarking Deep Reinforcement Learning for Continuous Control," *Proc. of the 33rd Int. Conf. on Machine Learning (ICML2016)*, Vol.48, Jun. 19-24, New York, pp. 1329-1338, 2016.
- [4] 美添一樹: "5分で分かる! ? 有名論文ナナメ読み: Silver, D. et al.: Mastering the Game of Go without Human Knowledge," 情報処理, Vol.59, No.8, pp. 750-752, 2018.
- [5] H. Igarashi and S. Ishihara: "An Algorithm of Policy Gradient Reinforcement Learning with a Fuzzy Controller in Policies," *Int. J. of Artificial Intelligence and Expert Systems*, Vol.4, pp. 17-26, 2013.
- [6] L. Joffe: "Fuzzy Inference System Learning by Reinforcement Methods," *IEEE Trans. on Systems, Man, and Cybernetics, Part C*, Vol.28, No.3, pp. 338-355, 1998.
- [7] 呉志賢, 中島智晴, 石淵久生: "ファジーQ学習の性能評価," 電子情報通信学会論文誌 D-II, Vol.J81-D-II, No.1, pp. 194-197, 1998.
- [8] 堀内匡, 藤野昭典, 片井修, 榎木哲夫: "連続値入出力を扱うファジィ内挿型Q-Learningの提案," 計測自動制御学会論文集, Vol.35, No.2, pp. 271-279, 1999.
- [9] 星野孝総, 亀井且有: "ファジィ環境評価ルールを用いた強化学習の提案とチェスへの応用," 日本ファジィ学会誌, Vol.13, No.6, pp. 626-632, 2001.
- [10] X. Wang, X. Xu, and H. He: "Policy Gradient Fuzzy Reinforcement Learning," *Proc. 3rd Int. Conf. on Machine Learning and Cybernetics*, pp. 992-995, 2004.
- [11] R. J. Williams: "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning," *Machine Learning*, Vol.8, pp. 229-256, 1992.
- [12] J. Baxter and P. L. Bartlett: "Infinite-Horizon Policy-Gradient Estimation," *J. of Artificial Intelligence Research*, Vol.15, pp. 319-350, 2001.
- [13] 五十嵐治一, 石原聖司, 木村昌臣: "非マルコフ決定過程における強化学習—特徴的適正度の統計的性質—," 電子情報通信学会論文誌 D, Vol.J90-D, No.9, pp. 2271-2280, 2007.
- [14] 杉本将也, 五十嵐治一, 石原聖司, 田中一基: "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: RoboCup 小型リーグにおける行動決定," 知能と情報, Vol.26, No.3, pp. 647-657, 2014.
- [15] 土屋義己, 五十嵐治一, 石原聖司: "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：自動車の速度制御におけるルール重みの学習,” 第 33 回ファジィシステムシンポジウム講演論文集, pp. 191-196, 2017.

[16] 関 宏理: “ファジィ関数型推論モデルの学習法,” 第 10 回情報科学技術フォーラム, pp. 175-180, 2011

[17] 関 宏理, 水本雅晴: “Type-2 ファジィ関数型推論モデル,” 第 27 回ファジィシステムシンポジウム講演論文集, pp. 76-81, 2011 (2019 年 8 月 3 日 受付) (2020 年 4 月 11 日 採録)

付録

式 (15) の導出

π' が式 (12) で与えられる場合, 特徴的適正度 $e'_\theta(t)$ は,

$$\begin{aligned} e'_\theta(t) &= \frac{\partial}{\partial \theta} \ln \pi'(y(t)) \\ &= -\frac{1}{T'} \left\{ \frac{\partial E'(y(t))}{\partial \theta} - \left\langle \frac{\partial E'(y)}{\partial \theta} \right\rangle_{\pi'(y)} \right\} \end{aligned} \quad (17)$$

と表される.

ここで, 式 (17) の右辺第 1 項について考える. 式 (11) より,

$$\frac{\partial E'(y(t))}{\partial \theta} = (y(t) - y_G(t)) \frac{\partial y_G(t)}{\partial \theta} \quad (18)$$

となる. また, 重心を θ で微分すると,

$$\begin{aligned} \frac{\partial y_G(t)}{\partial \theta} &= \frac{\partial \langle y(t) \rangle_{\pi(y)}}{\partial \theta} \\ &= \sum_y y \frac{\partial}{\partial \theta} \pi(y; x(t), \theta) \\ &= \sum_y \pi(y; x(t), \theta) y \frac{\partial}{\partial \theta} \ln \pi(y; x(t), \theta) \\ &= \langle y e_\theta(t) \rangle_{\pi(y)} \end{aligned} \quad (19)$$

となる. ただし,

$$\begin{aligned} e_\theta(t) &\equiv \frac{\partial}{\partial \theta} \ln \pi(y(t)) \\ &= -\frac{1}{T} \left\{ \frac{\partial E(y(t))}{\partial \theta} - \left\langle \frac{\partial E(y)}{\partial \theta} \right\rangle_{\pi(y)} \right\} \end{aligned} \quad (20)$$

である. 式 (20) を式 (19) へ代入すると,

$$\begin{aligned} \frac{\partial y_G(t)}{\partial \theta} &= -\frac{1}{T} \left\langle y \left\{ \frac{\partial E(y)}{\partial \theta} - \left\langle \frac{\partial E(y)}{\partial \theta} \right\rangle_{\pi(y)} \right\} \right\rangle_{\pi(y)} \\ &= -\frac{1}{T} \left\langle \left\langle y \frac{\partial E(y)}{\partial \theta} \right\rangle_{\pi(y)} - \langle y \rangle_{\pi(y)} \left\langle \frac{\partial E(y)}{\partial \theta} \right\rangle_{\pi(y)} \right\rangle_{\pi(y)} \\ &= -\frac{1}{T} \left\langle (y - y_G(t)) \frac{\partial E(y)}{\partial \theta} \right\rangle_{\pi(y)} \end{aligned} \quad (21)$$

と表される.

次に, 式 (17) の右辺第 2 項について考える. 式 (18) の両辺の期待値をとると,

$$\begin{aligned} \left\langle \frac{\partial E'(y)}{\partial \theta} \right\rangle_{\pi'(y)} &= \left\langle (y - y_G(t)) \frac{\partial y_G(t)}{\partial \theta} \right\rangle_{\pi'(y)} \\ &= \left\langle y \frac{\partial y_G(t)}{\partial \theta} \right\rangle_{\pi'(y)} - \left\langle y_G(t) \frac{\partial y_G(t)}{\partial \theta} \right\rangle_{\pi'(y)} \\ &= \left(\langle y \rangle_{\pi'(y)} - y_G(t) \right) \frac{\partial y_G(t)}{\partial \theta} \end{aligned} \quad (22)$$

となる.

式 (17) の右辺へ, 式 (18), 式 (22) を代入すると,

$$\begin{aligned} e'_\theta(t) &= -\frac{1}{T'} \left\{ (y(t) - y_G(t)) \frac{\partial y_G(t)}{\partial \theta} \right. \\ &\quad \left. - \left(\langle y \rangle_{\pi'(y)} - y_G(t) \right) \frac{\partial y_G(t)}{\partial \theta} \right\} \\ &= -\frac{1}{T'} \frac{\partial y_G(t)}{\partial \theta} (y(t) - \langle y \rangle_{\pi'(y)}) \end{aligned} \quad (23)$$

となる. さらに, 式 (21) を代入すると,

$$e'_\theta(t) = \frac{1}{TT'} \left\langle (y - y_G(t)) \frac{\partial E(y)}{\partial \theta} \right\rangle_{\pi(y)} (y(t) - \langle y \rangle_{\pi'(y)}) \quad (24)$$

となる. 一方, 式 (2) より,

$$\begin{aligned} \frac{\partial E(y; x(t), \theta, A, B)}{\partial \theta_i} &= \frac{\partial}{\partial \theta_i} \left[-\sum_{i=1}^{nR} \theta_i A^i(x(t)) B^i(y) \right] \\ &= -A^i(x(t)) B^i(y) \end{aligned} \quad (25)$$

である. 式 (24) と式 (25) から,

$$\begin{aligned} e'_\theta(t) &= -\frac{1}{TT'} \left\langle (y - y_G(t)) A^i(x(t)) B^i(y) \right\rangle_{\pi(y)} (y(t) - \langle y \rangle_{\pi'(y)}) \\ &= -\frac{A^i(x(t))}{TT'} \left\langle (y - y_G(t)) B^i(y) \right\rangle_{\pi(y)} (y(t) - \langle y \rangle_{\pi'(y)}) \end{aligned} \quad (26)$$

となり式 (15) を得る.

[問い合わせ先]

〒350-0394 埼玉県比企郡鳩山町石坂

東京電機大学理工学部

石原 聖司

E-mail: ishihara_s@mail.dendai.ac.jp

著者紹介

いしはら せいじ [非会員]
石原 聖司

2000 年法政大学大学院工学研究科システム工学専攻博士後期課程修了. 博士 (工学). 同年 (株) リコー入社. 2002 年近畿大学工学部講師, 2007 年同准教授. 2012 年東京電機大学理工学部准教授, 2016 年同教授, 現在に至る. 画像情報処理, 機械学習等の研究に従事. 電子情報通信学会, 電気学会, 情報処理学会, 画像電子学会などの会員.

いごらし ちかづ [非会員]
五十嵐 治一

1979 年東京大学教養学部基礎科学科卒業. 1981 年同大学院修士課程修了. 1984 年同大学理学系研究科博士課程単位取得退学. 同年 4 月 (株) リコー入社. 1990~1994 年の間, (株) ATR 視聴覚機構研究所と (株) ATR 人間情報通信研究所とへ出向. 1996 年 近畿大学工学部電子情報工学科助教授, 2002 年同教授. 2004 年芝浦工業大学工学部情報工学科教授, 現在に至る. 組合せ最適化問題, 自律移動型ロボットの走行制御, マルチエージェントシステム, 強化学習, コンピュータ将棋の研究に従事. 電子情報通信学会, 情報処理学会, 人工知能学会, 計測自動制御学会, 日本ロボット学会各会員. 博士 (理学).

Policy Gradient Reinforcement Learning for Policy Represented by Fuzzy Rules: Application to Simulations on Speed Control of an Automobile

by

Seiji ISHIHARA and Harukazu IGARASHI

Abstract:

A method of a fusion of fuzzy inference and policy gradient reinforcement learning has been proposed that directly learns, as maximizes the expected value of the reward per episode, parameters in a policy function represented by fuzzy rules with weights. A study has applied this method to a task of speed control of an automobile and has obtained correct policies, some of which control speed of the automobile appropriately but many others generate inappropriate vibration of speed. In general, the policy is not desirable that causes sudden time change or vibration in the output value, and there would be many cases where the policy giving smooth time change in the output value is desirable. In this paper, we propose a fusion method using the objective function, that introduces defuzzification with the center of gravity model weighted stochastically and a constraint term for smoothness of time change, as an improvement measure in order to suppress sudden change of the output value of the fuzzy controller. Then we show the learning rule in the fusion, and also consider the effect by reward functions on the fluctuation of the output value. As experimental results of an application of our method on speed control of an automobile, it was confirmed that the proposed method has the effect of suppressing the undesirable fluctuation in time-series of the output value. Moreover, it was also showed that the difference between reward functions might adversely affect the results of learning.

Keywords: fuzzy controller, reinforcement learning, policy gradient method, center of gravity model, speed control problem

Contact Address: **Seiji ISHIHARA**

*School of Science and Engineering, Tokyo Denki University
Ishizaka, Hatoyama-machi, Hiki-gun, Saitama 350-0394, Japan
E-mail: ishihara_s@mail.dendai.ac.jp*