2D-7

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の提案

五十嵐 治一† 石原 聖司‡
芝浦工大† 近畿大‡

1. はじめに

ファジィ推論と強化学習の組合せは以前から多く試みられてきた[1][2]．ファジィ推論の立場からは，if-then ルール内の各種パラメータの値を強化学習により自動調整ができ，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のためにファジィ集合が利用できる．

しかし従来は，Q 学習のような MDP を仮定した価値ベースの強化学習を用い，ルールの後件部が関数や定数で表現されている場合が殆どであった．さらに，複数ルールの出力値の統合結果を非ファジィ化する際には重心計算を用いる例が多いが，システム制御への応用等では不都合が生じる場合もあった．このような背景の下に，本研究では一つの融合方式を提案する．

2. ファジィルールとファジイ推論

2.1 ファジィルール

本稿ではファジィ集合を含む if-then 型ルール（ファジィルール）を考える．以下に示すのは自動車の運転制御についての例である[2].

規則１：車間距離が小さく，速度が速いならばブレーキ力を強くする． (1)
規則２：車間距離が大きく，速度が遅いならばブレーキ力を弱くする． (2)

ルール前件部の「小さい／大きい」「速い／遅い」や，後件部の「強く／弱く」などの程度を表す言葉はファジィ集合で表現される．

2.2 ファジィ推論：Min-Max 重心モデル

代表的なファジィ推論モデルである Min-Max 重心モデル（Mamdani の推論モデル）を説明する．(1), (2)の例において，車間距離, 速度からなる入力変数を $x=(x_1, x_2)$，ブレーキ力を表す出力変数を y とする．また，車間距離が「小さい」「大きい」を表すファジィ集合（のメンバシップ関数）を A_1^1, A_1^2, 速度が「速い」「遅い」を A_2^1, A_2^2, ブレーキ力を「強く」「弱く」を B^1, B^2 で表すと，(1),(2)の 2 つの規則は,

Rule 1: If x_1 is A_1^1 and x_2 is A_2^1 Then y is B^1 (3)
Rule 2: If x_1 is A_1^2 and x_2 is A_2^2 Then y is B^2 (4)

と表される．Rule i (i=1,2)における推論結果 B^{i*}は，前件部の適合度 $\min\{A_1^i(x_1), A_2^i(x_2)\}$と，後件部の適合度 $B^i(y)$との Min 演算により，

$$B^{i*}(y) = \min\{ \min\{A_1^i(x_1), A_2^i(x_2)\}, B^i(y) \} \quad (5)$$

で与えられる．次に，２つの規則の推論結果を和集合 $B^*= B^{1*} \cup B^{2*}$により統合する（メンバシップ関数の Max 演算）．さらに，結論である出力変数の代表値 y^*を次の重心計算により得る．

$$y^* = \int yB^*(y)dy \Big/ \int B^*(y)dy \quad (6)$$

他の代表的なファジィ推論法である「T-S 推論モデル」は，後件部のファジィ集合を関数（多くは一次式）に置き換えたモデルであり，さらに定数で置き換えたモデルは「簡略化ファジィ推論モデル」と呼ばれ，実応用で広く用いられている．しかし，いずれも複数規則の推論結果の統合して非ファジィ化する際に，(6)のような重心計算により出力値を決定論的に決めている．重心計算は規則ごとの結論に重なりが少ない場合，例えば Mamdani の推論モデルで(6)の和集合 B^*のメンバシップ関数に多峰性がある場合には，適合度の低い y の値が最終結論とされてしまう．

3. 方策勾配法

方策勾配法は強化学習の一手法である[3]．Q 学習などの価値関数の計算にベースを置く手法ではなく，方策関数中のパラメータを学習することにより方策を直接的に構築する．if-then 型の状態行動ルールや，ポテンシャルなどの形式を持つヒューリスティクスを方策へ反映させやすく，また，MDP の仮定も必要としないので現実世界への広い応用が期待できる[4].

時刻 t，状態 s，行動 a が離散的であり，目的関数 $E(a(t);s(t),h_t,\omega)$を用いた Boltzmann 分布によるエージェントの確率的方策

$$\pi\left(a(t);s(t),h_t,\omega\right) \equiv \frac{e^{-E\left(a(t);s(t),h_t,\omega\right)/T}}{\sum_{a\in A} e^{-E\left(a;s(t),h_t,\omega\right)/T}} \quad (7)$$

を考える．h_t は状態と行動の履歴，ωは学習パラメータである．方策勾配法によるとエピソードごとの報酬期待値を極大化する学習則は,

$$\Delta\omega = \varepsilon \cdot r \cdot \sum_{t=0}^{L-1} e_\omega(t) \quad (8)$$

で与えられる[3][4]．ただし，L はエピソード長,

Policy Gradient Reinforcement Learning with a Fuzzy controller for Policy
†Harukazu Igarashi, Shibaura Inst. of Tech.
‡Seiji Ishihara, Kinki Univ.

rは報酬，εは学習係数，$e_\omega(t)$は特徴的適正度

$$e_\omega(t) \equiv \partial \ln \pi\left(a(t);s(t),h_t,\omega\right)\Big/\partial\omega \qquad (9)$$

である．ωの更新はエピソード終了時ごとに行う．

4. ファジィ制御ルールへの拡張

4.1 基本方針

ファジィ推論と Q 学習とを組み合わせた学習法として，ファジィ Q 学習をはじめ様々な方法が考案されている[1][5]．そこでは後件部に状態や行動の価値を関数や定数等の非ファジィ集合で表現し，後件部内のパラメータを強化学習で学習する場合が多い．

本研究ではこれらを考慮した上で，①ファジィ集合による前／後件部の表現，②確率的な出力値の決定法，③メンバシップ関数の学習機能，④ルール重みの導入，⑤方策勾配法による学習機能の５つの特徴を持つ推論方式を提案する．このうち，②は重心計算による非ファジィ化の弊害の緩和であり，④と⑤は将来的にルールの生成・削除を学習させるためである．

上記５つの機能を実現するために，本手法では，ルール自体の重み θ_i ($\geqq 0$, i=1,2,…, n_R)を持った次のファジィルールを用いる：

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 (10)

4.2 目的関数

(7)の目的関数として以下の関数を用いる．

$$E(y;x,\theta,A,B) = -\sum_{i=1}^{N} \theta_i A^i(x) B^i(y) \qquad (11)$$

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

$$A^i(x) \equiv \prod_{j=1}^{M} A_j^i(x_j),\ B^i(y) \equiv \prod_{j=1}^{N} B_j^i(y_j) \qquad (12)$$

で定義する．

目的関数 $E(y;x)$は，入力値が x のときに出力値 y が適切であるかの度合いを表す．(11)ではルールごとに，前/後件部の適合度，ルール重みの積で y の適切さを表現し，それらの和を全ルールの統合結果とする．さらに y を重心計算ではなく，(7)の Boltzmann 分布により確率的に決定する．

4.3 学習則

(11)の目的関数を持った(7)の方策 π から，学習則(8)と特徴的適正度(9)の具体的な表式を導出した．前件部/後件部の各メンバシップ関数とルール重みの学習則は以下のようにまとめられる．

$$\Delta A^i{}_j(x) = \varepsilon \cdot r \cdot \sum_{t=0}^{L-1} e_{A^i{}_j(x)}(t) \qquad (13)$$

$$e_{A^i{}_j(x)}(t) = \theta_i \cdot \delta_{x(t),x} \cdot \prod_{l\neq j} A_l^i(x)\left[B^i(y(t)) - \langle B^i\rangle(t)\right]\Big/T \qquad (14)$$

$$\left[\begin{array}{l} \langle B^i\rangle(t) \equiv \sum_y B^i(y)\pi(y;x(t),\theta,A,B) \\ \delta_{x,x'} \equiv 1 \ \text{if}\ x = x' \ \text{else}\ 0 \end{array}\right] \qquad (15)$$

$$\Delta B_j^i(y) = \varepsilon \cdot r \cdot \sum_{t=0}^{L-1} e_{B_j^i(y)}(t) \qquad (16)$$

$$e_{B_j^i(y)}(t) = \theta_i \cdot A^i(x(t)) \cdot\left[\delta_{y(t),y} - \pi(y;x(t),\theta,A,B)\right]\prod_{l\neq j} B_l^i(y)\Big/T \qquad (17)$$

$$\Delta\theta_i = \varepsilon \cdot r \cdot \sum_{t=0}^{L-1} e_{\theta_i}(t) \qquad (18)$$

$$e_{\theta_i}(t) = A^i(x(t))\left[B^i(y(t)) - \langle B^i\rangle(t)\right]\Big/T \qquad (19)$$

上記の学習則では，各ルールとも出現した入/出力変数値のメンバシップ値だけが強化され，その強化量は，報酬値，前/後件部の適合度，ルール重みに応じて大きくなる．ルール重みは，前/後件部の適合度の積に比例して強化される．

5. 今後の展開

今回は入/出力変数の値に対する前/後件部のメンバシップ関数の値を学習パラメータとしたが，連続変数は取り扱うことができない．今後は，メンバシップ関数をパラメータ付きの連続関数で表現し，そのパラメータの学習則を導出する．また，本学習方式は多段ファジィ推論へそのまま拡張できる．さらに，階層型制御システムやマルチエージェントシステムへの応用も可能であり，応用を試みる予定である．

文献

[1] Jouffe, L., “Fuzzy Inference System Learning by Reinforcement Methods”, IEEE Transactions on Systems, Man, and Cybernetics, Vol.28, No.3, pp.338-355 (1998).

[2] 関，水本，“ファジー理論の現状と最近の動向”，電子情報通信学会誌 Vol.94, No.10, pp.908-914 (2011).

[3] Williams, R.J., “Simple Statistical Gradient- Following Algorithms for Connectionist Reinforcement Learning,” Machine Learning, vol.8, pp.229-256 (1992).

[4] 石原，五十嵐，”マルチエージェント系における行動学習への方策こう配法の適用-追跡問題-“, 電子情報通信学会論文誌 D-I, Vol.J87-D1, No.3, pp.390-397 (2004).

[5] 堀内，藤野，片井，椹木，” 連続値入出力を扱うファジイ内挿型 Q-Learning の提案 “，計測自動制御学会論文集，Vol.35, No.2, pp.271-279 (1999).