FEERL(Fuzzy Environment Evaluation Reinforcement Learning) を用いた習得済みルールの有効利用

Effective Reuse of Acquired Rules by FEERL

○星野 孝総　　亀井 且有
Yukinobu Hoshino　　Katsuari Kamei

立命館大学理工学部　　立命館大学理工学部
Computer Science　　Computer Science
Ritsumeikan University　　Ritsumeikan University

Abstract Reinforcement learning is a powerful machine learning system, which is able to learn without giving training examples to a learning unit. But it is impossible for the reinforcement learning to support large environments because the number of *if-then* rules is a huge combination of a relationship between one environment and one action. We propose new reinforcement learning system ,FEERL ,for the large environment. In this paper, we have tried to use acquired rules on maze problem.

1　はじめに

熟練者の教師データがない場合，エージェントは試行錯誤によって学習を進めなければならない．未知環境に対する学習手法として，強化学習が提案されている．代表的な手法である Q-learning[7][8] は，状態と行動を対にしたルールを Q 値で評価し，政策・行動を決定する．したがって，ルールを類似状態や異種類の行動有する環境に利用できない．さまざまな環境に対し，ルールを有効利用する手法を強化学習に組み込む必要がある．筆者らは，ファジィ環境評価型強化学習手法 (FEERL) を文献 [4][5][6] で提案し，チェスゲームを用いて，複雑で巨大な環境に対する有効性を示した．本論文では，迷路探索で習得したルールを，行動形態変更し類似迷路を有した迷路探索問題に適用し，その有効性を検証する。

2　強化学習

強化学習は状態認識器，行動選択器，学習器の三つのユニットから構成されている．状態認識器は状態を認識し，政策候補の集合を生成し，行動選択器に送る．行動選択器は，状態認識器から送られた政策候補の集合から評価値の大きい行動を選択して環境に出力する．この政策により状態が遷移し，遷移先状態が報酬・罰の条件を満たしているとき環境は報酬・罰を学習器に与える．学習器は，報酬・罰に従って政策に関する評価値を変更する．強化学習での，報酬 (reward)，罰 (penalty) は政策に対して遅れがあり，得られる条件は遷移先状態によって決定される．したがって，学習の目的はその報酬を多く得ることであり，言いかえれば，時間軸上の未来に対する報酬の総和を最大にすることである [1][2].

3　ファジィ環境評価型強化学習 (FEERL)

環境評価に基づく強化学習手法として TD 法 [9] があるが，巨大環境に適用する場合，感覚入力を評価関数により変換し, あつかえる程度の次元数で評価しなければならない。したがって，TD 法による学習は評価関数に大きく依存する．FEERL では，感覚入力を過去のルールを用い，ファジィ推論により直接評価できる．この点が TD 法と大きく異なる．

3.1　環境シミュレータ

状態評価を用いる場合，行動を生成する機構が必要となる．FEERL は，環境と行動から遷移先環境を生成できる機構を採用している．環境シミュレータは環境と行動を入力とし，遷移先環境を出力とするシステムである．ただし，遷移先環境を出力できない行動ベクトルが入力された場合，その行動はエージェントがとれない行動と考える．

3.2 ファジィ環境評価ルール

環境評価型強化学習における環境 $\boldsymbol{q}$ を N 次元ベクトルで与えるとする．

$$\boldsymbol{q} = (q_1, q_2, \cdots, q_j, \cdots, q_N) \tag{1}$$

また，環境評価型強化学習における環境評価ルール R_i を (1) 式で与える．ただし， $\boldsymbol{p}_i$ は，過去に出現した環境，E は環境評価値を示す．(1) 式は i 番めのルール R_i において，入力環境が $\boldsymbol{p}_i$ のとき，環境評価値が w_i であることを表している．

$$\begin{aligned} R_i: \quad & if \boldsymbol{q}\ is\ \boldsymbol{p}_i then\ E\ is\ w_i \\ & \boldsymbol{p}_i = (p_{1i}, p_{2i}, \cdots, p_{ji}, \cdots, p_{Ni}) \\ & i\text{=}1,2,\cdots,M;\ M\ \text{はルール数} \end{aligned} \tag{2}$$

まず，任意な環境 $\boldsymbol{q}$ と環境評価ルール R_i の前件部 $\boldsymbol{p}_i$ から，三角型メンバシップ関数により適合度 μ_{ji} を (3) 式により算出する．

$$\mu_{ji} = \begin{cases} 1 - \dfrac{|p_{ji} - q_j|}{l} & |p_{ji} - q_j| < l \\ 0 & |p_{ji} - q_j| \geq l \end{cases} \tag{3}$$

ここで， $\boldsymbol{q}$ と $\boldsymbol{p}_i$ の類似度 $\widehat{\mu}_i$ を (3) 式で定義する．m_i は全次元に対する適合度であり，L は $\widehat{\mu}_i$ の不感帯を決定するパラメータ，pp は の感度パラメータである．たとえば，L を小さく設定すれば，m_i がかなり大きくても類似度は零となり，pp を大きく設定すればわずかな m_i の変化で $\widehat{\mu}_i$ が大きく変化する．

また，一般的なファジィ推論での類似度の定義は，$\widehat{\mu}_i = \bigwedge_j \mu_{ji}$ で与えられるが，これでは多次元中の一次元でも μ_{ji} が小さければ，全く類似しない状態と判断される．環境 $\boldsymbol{q}$ における遷移先環境評価値 E($\boldsymbol{q}$) を (10) 式にように類似度 $\widehat{\mu}_i$ を重みとする環境評価値 w_i の重み付き平均で与える．このように，環境評価ルールが発火しない遷移先環境 (未知環境) が入力された場合でも，FEERL は遷移先環境評価値 E($\boldsymbol{q}$) を算出できる．

$$\widehat{\mu}_i = \begin{cases} \left(1 - \dfrac{m_i}{L}\right)^{pp} & L > m_i \\ 0 & L \leq m_i \end{cases} \tag{4}$$

ただし，$m_i = \sum_{j=1}^{N} \mu_{ji} \quad , pp \geq 1 \quad , 0 < L < N$

$$E(\boldsymbol{q}) = \begin{cases} \dfrac{\sum_{i=0}^{M} w_i \widehat{\mu}_i}{\sum_{i=0}^{M} \widehat{\mu}_i} & \sum_{i=0}^{M} \widehat{\mu}_i > 0 \\ 0 & \sum_{i=0}^{M} \widehat{\mu}_i = 0 \end{cases} \tag{5}$$

3.3 探索アルゴリズム

行動を決定できない環境において，エージェントは限られた環境評価ルールにもとづいて行動を決定しなければならない．FEERL は，環境シミュレータが算出した遷移先環境とファジィ環境評価ルールによる遷移先環境評価値 $E(\boldsymbol{q})$ から先読み木を生成し，ゲーム理論の探索アルゴリズムを用いて行動選択を行う．

3.4 評価値の強化

ファジィ環境評価ルールの環境評価値は，報酬と環境同定により強化され，(6) 式によって与えられる．

$$w_i \longleftarrow (1 - \alpha) w_i + \alpha(r + \gamma E^{max}) \tag{6}$$

α は学習率である．また，r は環境に対する期待報酬である．E^{max} は行動選択された行動に対する環境シミュレータにより木探索された深度分先の遷移先環境評価値である．これは行動を起こした時の未来における報酬の見積もり値と考えることができる．また，γ は割引率であり，行動連鎖のつながりを示している．α と γ は，それぞれ [0,1] の値をとる．FEERL は，ファジィ推論で用いた環境評価ルールを強化し，その報酬は環境の状態のみで判断される．また，環境評価ルール R_i の後件部 w_i は各行動ごとに強化されるので，過去に報酬を受けない環境においても，ルール R_i の類似度 $\widehat{\mu}_i$ および E^{max} から環境評価値 w_i を強化することができる．

4 部分観測迷路問題への適用

図 1 に示すような環境を用い，提案手法を迷路探索問題に適用した．エージェントは，5×5 の視界を持っている．視界が小さいためファジィ環境評価ルールを使用する必要がない．そこで使用するルールは，エージェントが観測した状態とその評価値とし，また環境シミュレータは状態, 行動，遷移先状態を記憶させてリアルタイムで作成した．先読み深度は，3 ステップとした．エージェントは，スタート (S) を出発し，ゴール (G) に到達

したときに報酬を与え，1 試行が終了する．25 ステップ毎に罰を与え，学習はステップ毎に強化した．したがってほとんどの学習は，報酬 $r = 0$ でおこなう．

4.1 迷路探索実験

実験は，66 試行おこなった，結果を図 2 に示す．グラフは，試行回数とゴールまでに要したステップ数を示している．試行回数に対して，ステップ数が減少している事が解る．また，図 1 に獲得した行動を実線矢印で示す．

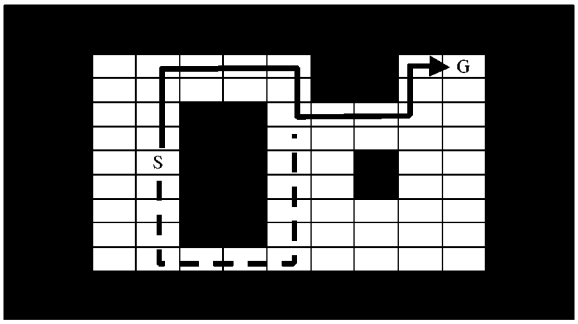

図 1 Maze and Best Route

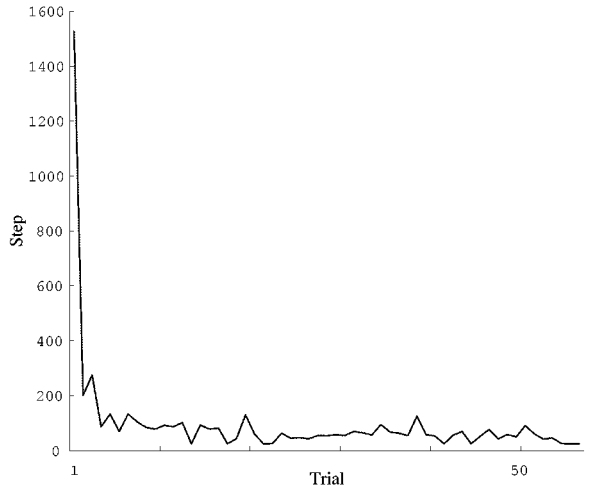

図 2 Relationship between Trial and Step

5 FEERL による巨大環境への習得済ルール有効利用

迷路探索実験で取得したルールを，類似状態や異種類の行動を有する迷路問題に適用する．環境は，迷路探索実験で学習した迷路の大きさを縦・横に 10 倍とし，図 3 のように複雑な形状にした迷路を用意する．迷路は 52×52 であり，エージェントの視界は 20×20 とする．したがって，状態パラメータの次元数は 400 となる先読み深度は，3 ステップとした．エージェントは，スタート (S) を出発し，ゴール (G) に到達したときに報酬を与え，1 試行が終了する．40 ステップ毎に罰を与え，学習はステップ毎に強化した．したがってほとんどの学習は，報酬 $r = 0$ でおこなう．エージェントは，迷路内を (7) 式に示すような，運動方程式にしたがって移動するとし，行動は速さ v とハンドル角度 θ の操作量とする．ここで，x, y は縦・横位置，$\dot{x}, \dot{y}$ は縦・横移動量，ϕ は車体の角度，$\dot{\phi}$ は車体の角速度，L は車体のホイールベースである．迷路探索実験で習得したルールを縦・横に 4 倍に変更し，新しいルールの初期とする．

$$\begin{cases} L = 0.5 \\ \phi = \phi + \dot{\phi} \\ x = x + \dot{x} \\ y = y + \dot{y} \\ \dot{x} = cos(\phi) \times v \\ \dot{y} = sin(\phi) \times v \\ \dot{\phi} = tan(\phi) \times v \div L \end{cases} \tag{7}$$

エージェントは，視界情報のみを入力とし，速さ $v = (1.0, 0.8, 0.6, 0.2, -0.2)[m]$ とハンドルの操舵角度 $\phi = (-60, -30, 0, +30, +60)[deg]$ を出力とする．最初の格子空間で習得した環境評価ルールを初期値に用いる．エージェントの視界は 16 倍になっており，状態の組合せが大きくなる．したがって，ファジィ環境評価ルールを用いる．

格子空間の迷路では，環境シミュレータに状態パラメータと行動履歴データとのパターンマッチングを用いて遷移先状態をデータから選択した．実験では，過去 2 行動の履歴，候補行動からデータの状態パラメータと遷移先状態の差を推論するモデルを作成し用いた．推論に用いたメンバーシップの広がりは，$l = 0.5$,$L = 3$,$PP = 3$ とした．

次に，推測された遷移先状態をファジィ評価値ルールを用いて評価する．変換されたルールでは，4×4 のエリアにルールが影響していると考える．そこで，その部分がグレード値 0.5 で交わるように縦横 8 マスの四角錐型メンバシップ関数を設定する．実験では，各マスの適合度を四角錐型メンバシップ関数から計算し，各マスで得られた適合度からファジィ類似を計算し評価値を算出する．強化学習のパラメータは $\alpha = 0.2$，$d = 0.9$ とした．また，メ

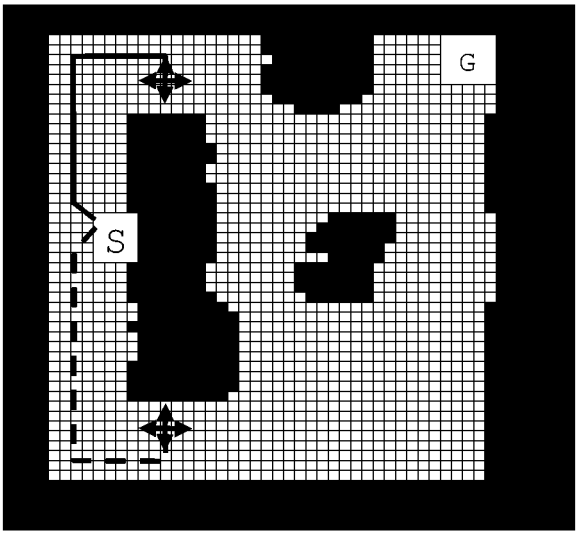

図 3 Car Type Env.

ンバーシップの広がりをそれぞれ $l = 8$,$L = 280$,$PP = 3$ とした．

実験では，13 試行で図 3 の矢印が示す経路を示した．初期ルールの影響で壁に沿って行動している．しかしながら，矢印の先で迷ってしまう現象が確認された．この地点での遷移先状態を計算するデータが少ないため，正確な評価ができなくなり，ランダム探索になっているのではないかと考えられる．また，破線で示された経路も多く観測された．これは，図 1 の破線経路が学習されていたため，変換したルールにそれらが反映されていたためであると考えられる．実験は 34 試行まで行った．しかし，最後までエージェントはスタート付近のランダム探索を続けていた．

6 おわりに

本論文では，簡単な部分観測迷路探索問題を環境評価型強化学習に解かせ，習得したルールを 16 倍もの大きい類似環境に適用し，その有効性を検証した．状態爆発を起こす環境において，本手法を用いたルールの有効利用を確認することはできなかった．また，大きい環境での学習は収束せず．反対に学習アルゴリズムが行動を混乱させる結果になった．これは，先読みアルゴリズムに使用した環境シミュレータの性能が良くないためであると考えられる．また，実験で学習されたルールを解析し，最適なパラメータを検証する必要がある．そこで今度の課題として，精度の高い環境シミュレータ構築，最適パラメータの検証，連続値出力へのアプローチを試みる．

参考文献

[1] 畝見達夫：強化学習；人工知能学会誌，Vol.9，No.6，pp.40-46 (1994)

[2] 畝見達夫：実例に基づく強化学習法；人工知能学会誌, Vol.7,No.1, pp.141-151 (1992)

[3] 堀内，藤野，片井，椹木: 連続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)

[4] 星野, 亀井: 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; 日本ファジィ学会第 10 回ソフトサイエンスワークショップ予稿集, pp.98-101 (2000)

[5] Yukinobu Hoshino, Katsuari Kamei: A Proposal of Learning System with Fuzzy Rules Under Large Environments; Processing Of International Joint Conference on Neural Network 99 (1999)

[6] Y.Hoshino and K.Kamei: A proposal of Learning System with Fuzzy Rules under Large Environments Proceedings of Asian Fuzzy System Symposium, Tsukuba Science City, Japan, pp.179-184, (2000)

[7] Watkins.C.J.C.H: Learning from delayed rewards; Doctoral thesis, Cambridge University, Cambridge, England (1989)

[8] Watkins.C.J.C.H., Daya.P: Technical Note: Q-Learning; Machine Learning, Vol.8, No.3, pp.279-292 (1992)

[9] Richard S.Sutton: Learning to Predict the Methods of Temporal Differences; Machine Learning Vol.3, No.2/3, pp.9-44 (1988)

[**お問い合わせ**]
星野孝総
立命館大学理工学部
滋賀県草津市野路東 1-1-1