# ファジィ環境評価型強化学習 (FEERL) を用いた知識の有効利用

# Effective use of Learning Knowledge by Fuzzy Environment Evaluation Reinforcement Learning (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** The machine learning is proposed to learn techniques of specialists. A machine has to learn techniques by trial and error when there are no training examples. 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 have proposed new reinforcement learning system for the large environment, Fuzzy Environment Evaluation Reinforcement Learning (FEERL). In this paper, we proposed to reuse of the acquired rules by FEERL.

## 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．代表的手法である Q-learning[5, 6] を用いた場合，習得した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する場合，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を文献 [3][4] 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有効利用できることを示す．

## 2　強化学習

強化学習 [1, 7] 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．

## 3　ファジィ環境評価型強化学習

ファジィ環境評価型強化学習 (以下 FEERL と示す) は，ファジィ推論を用いた強化学習 [2] の一手法である．この手法では，状態を評価するためのルールベースを過去の経験としてシステム内部に持ち，未知状態に対しファジィ推論によって状態評価を行う．この評価値と先読みアルゴリズムにより未来の状態評価を行いながら学習・探索を行うことができる．したがって，過去に経験した状態から未知状態に対する行動決定でき，学習させることが可能な機械学習アルゴリズムである．

### 3.1　ファジィ類似度推論法

FEERL では，入力状態 $\boldsymbol{q}$ と環境評価ルールの前件部 $\boldsymbol{p}_i$ との類似度 $\widehat{\mu}_i$ を (2) 式で与える．ここで，各要素 $p_{ji}$ の適合度を $\mu_{ji}$ とする．図 1 に示すように，$m_i$ は全次元に対する適合度の総和であり，$L$ は $\widehat{\mu}_i$ の不感帯を決定するパラメータ，$pp$ は感度パラメータである．この

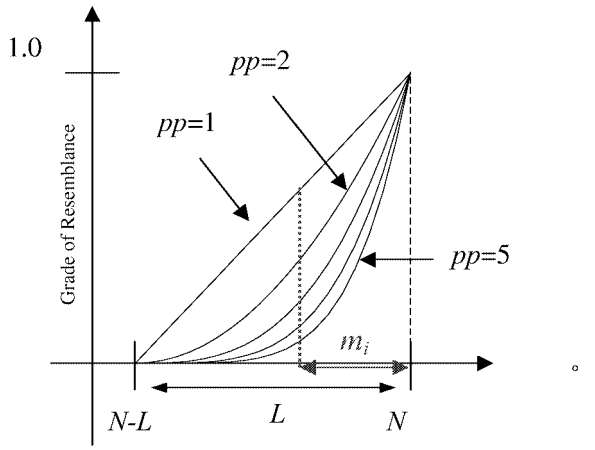


図 1 不感帯をモデルとしたファジィ環境類似度推論

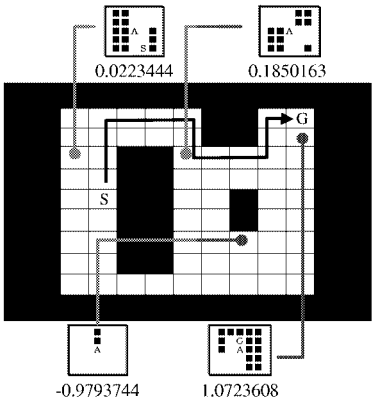


図 2 習得したルールの例と獲得した経路

ような，ファジィ類似度推論により，入力状態 $\boldsymbol{q}$ における環境評価値 $E(\boldsymbol{q})$ を (3) 式にように類似度 $\widehat{\mu}_i$ を重みとする環境評価ルールの後件部環境評価値 $w_i$ の重み付き平均で与える．ただし，$m_i = N - \sum_{j=1}^{N} \mu_{ji}$, $pp \geq 1$, $L = \lambda\ N$ とする．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．しかし，これでは多次元中の一次元でも $\mu_{ji}$ が小さければ，全く類似しない状態と判断される．このような判断は人間が行う状態評価の感覚と一致しない場合がある．

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

$$\widehat{\mu}_i = \begin{cases} \left(1 - \dfrac{m_i}{L}\right)^{pp} & m_i < L \\ 0 & m_i \geq L \end{cases} \tag{2}$$

$$E(\boldsymbol{q}) = \begin{cases} \dfrac{\sum_{i=0}^{M} w_i \widehat{\mu}_i}{\sum_{i=0}^{M} \widehat{\mu}_i} & \sum_{i=0}^{M} \widehat{\mu}_i > 0 \\ 0 & \sum_{i=0}^{M} \widehat{\mu}_i = 0 \end{cases} \tag{3}$$

## 4 部分観測迷路問題への適用

図 2 に示すような環境を用い，環境評価ルール [3] を用いた環境評価型強化学習を迷路探索問題に適用した．この時，環境評価型強化学習の環境シミュレータは，実際に体験した状態とその時の行動から構築するモデル構築型を用いた．具体的には，過去 3 ステップの行動履歴，観測状態および候補行動から遷移先状態を出力するデータベースを作成し，行動毎に検索した．また，見つからない場合その候補行動は行動不可能とし，先読みを枝刈りする．したがって，先読みアルゴリズムでの遷移先状態の評価値は 0 とした．エージェントは，5×5 の視界を持ち，観測状態に対し，その評価値を基に行動を決定する．また，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．

実験は，66 試行おこなった，結果を図 3 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 2 に獲得した行動を実線矢印で示す．また，図中 4 角の図と値は，習得したエージェントの代表的な視界状態と評価値であり，中央部の迷路に対応している．

## 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 4 倍とし，図 6 のように複雑な形状を持つ迷路を用いる．迷路は $52 \times 52$ であり，エージェントの視界は $20 \times 20$ とする．先読み深度は，3 ステップとした．先読みは深度 5，ファ

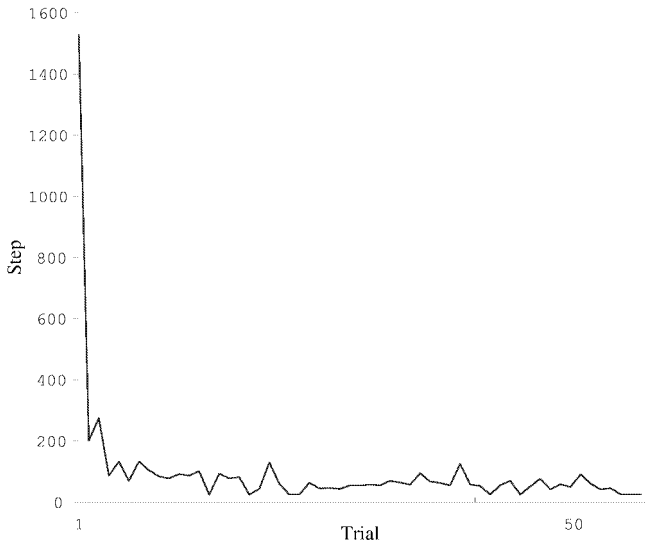


図 3 ゴール到達までのステップ数

ジィ類似度推論法のパラメータは，$l = 2, \lambda = 0.5, pp = 5, N = 20 \times 20 = 400$ とした．エージェントは，スタート (S) を出発し，ゴール (G) に到達する事を目的とする．迷路探索実験で習得したルールを図 4 に示すように縦・横 4 倍とし，ルールベースの初期とする．エージェントは，視界情報のみを入力とし，図 5 に示すように前進 $(-1.0, 0.0, 1.0, 2.0)$，回転 $(-0.5, 0, +0.5)[deg]$ を出力とした．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．また，エージェントの視界は 16 倍になる． 実験結果を図 6 に示す．エージェントは，直進と回転による方向転換を繰り返し，ゴールたどり着く経路が観測された．図 7 上部に示すように習得済みルールはサブゴールの役目を果たし，先読みアルゴリズムがサブゴール付近を探索している考えられる．従って，先読み探索の範囲にサブゴールがあれば，間の状態に無関係にゴールにたどり着くことができ，言い換えれば観測状態と探索先状態との間にマルコフ性がある事になる．この状態でゴールに到達することで，エージェントはより細かい状態を体験しながら目的を達成できる事になる．このような状態に対し，図 7 下部に示すような追加学習を FEERL で行えば，更に高度なルールが生成できると考えられる．また，FEERL でファジィ類似度推論で評価できる程度にルールを生成できれば，先読み探索の深度を小さく出来ることが可能になり，より高速に行動選択を行うことが出来るようになる．

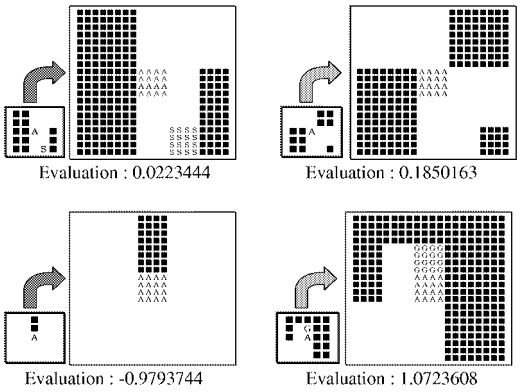


図 4 ルール視界部の拡張

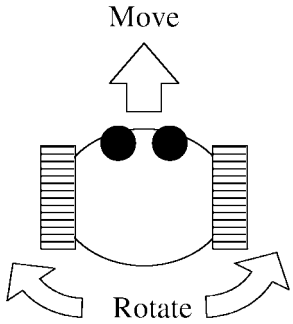


図 5 エージェントの行動

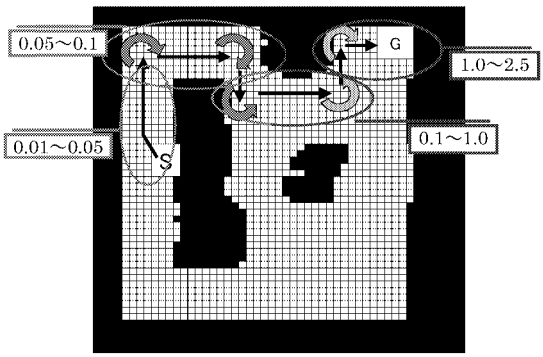


図 6 巨大な類似環境と実験結果

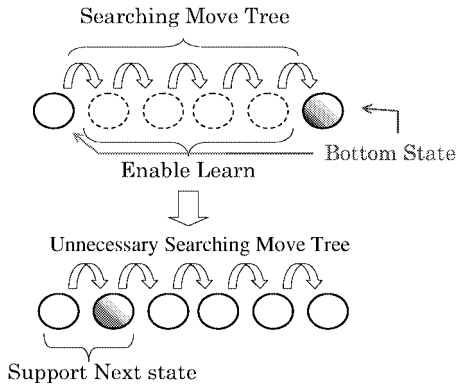


図 7 追加学習による効果

## 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 倍大きい類似環境に適用した．結果，習得したルールはファジィ類似度推論法を用いる事で，類似環境に有効利用可能である事が解った．この状態を FEERL で追加学習すれば，より高品質なルールを取得できる．先読み経路上のルールを習得できれば，ファジィ推論を用いて評価値を補間することが可能になるため，先読みアルゴリズムの深度を小さくできると考えられる．今後は，推論法の改良を行うとともに，FEERL による追加学習と深度パラメータの関係を調査していきたい．また，モデル構築型 FEERL の改良もおこなう予定である．

## 参考文献

[1] 畝見：強化学習；人工知能学会誌，Vol.9，No.6，pp.830–836 (1994)

[2] 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)

[3] 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; 日本ファジィ学会誌，Vol.13，No.6，pp.626–632 (2001)

[4] 星野，亀井: ファジィ環境評価型強化学習の LightsOut ゲームへの応用と探索における迂回行動の回避システム制御情報学会 論文誌，Vol.14，No.8，pp.395–401 (2001)

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

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

[7] R.S.Sutton and A.G.Barto：Reinforcement Learning；*The MIT Press*

**[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