不応性を有する領域表現を用いたKFM連想メモリによる強化学習の実現

清水厚志　長名優子

東京工科大学 コンピュータサイエンス学部

1 はじめに

環境との相互作用により適切な行動系列を獲得するための学習手法として，強化学習に関する様々な研究が行われている．強化学習では，設計者が意図しない未知の環境やノイズの多い実環境においても学習が行えるという特徴がある [1].

本研究では，強化学習の一種であるアクター・クリティック [2] のアクターの部分に不応性を有する領域表現を用いたKFM(Kohonen Feature Map) 連想メモリ [3] を用いた強化学習を提案する．不応性を有する領域表現を用いたKFM連想メモリは逐次学習が可能なモデルであり，これを強化学習に用いた場合，タスクを実行しながら学習を行うことができる．また，環境の変動などに対しても柔軟に対応できると考えられる．

2 強化学習

2.1 強化学習とは

強化学習は，エージェントとエージェントがおかれた環境との相互作用によって進行する．エージェントは環境から状態 s_t を観測し，s_t に基づいて行動 a_t をおこし，その結果，状態 s_{t+1} に遷移するとともに報酬 r_t を得るというサイクルを繰り返しながら学習を進めていく．なお，状態から行動を導き出す規則は方策と呼ばれる．

2.2 アクター・クリティック

アクター・クリティックは，価値の推定を行う評価部分 (クリティック) と，行動を選択するために用いられる行動部分 (アクター) とから構成されており，クリティックにおいてTD誤差を出力することから，TD学習の一種としてみることができる．

クリティックは状態価値関数に対応し，各行動の後で新しい状態を評価し，実行結果が期待されたものより良かったかどうかを判断する．その評価結果がTD誤差となる．アクターでは観測された状態とクリティックから出力されたTD誤差をもとに確率的に行動選択を行う．

3 不応性を有する領域表現を用いたKFM連想メモリによる強化学習の実現

提案手法では，アクター・クリティック [2] のアクターの部分を不応性を有する領域表現を用いたKFM連想メモリ [3] を用いて実現する．不応性を有する領域表現を用いたKFM連想メモリの入出力層を状態と行動を表す部分に分け，状態を入力としたときに，行動を出力できるように学習を行う．

不応性を有する領域表現を用いたKFM連想メモリは逐次学習が可能なモデルであるため，はじめにすべての状態と行動の関係を学習しておく必要はなく，タスク中にも学習を行うことができる．また，不応性を有するモデルであるため，1つの状態から複数の行動を出力することができる．これらの特徴は，強化学習を行う際，環境に変動があるような場合に有効に働くと考えられる．

また，クリティックでは環境から得られる状態を入力とし，価値の更新や評価を行う．さらに，アクターへTD誤差を出力する．アクターとして動作する不応性を用いた領域表現を用いたKFM連想メモリ (以下，アクターネットワークと呼ぶ) では，クリティックから出力されたTD誤差に基づいて学習を行い，環境の状態から行動の選択を行う．

(1) アクターネットワークの重みを小さなランダムな値で初期化する．

(2) エージェントが環境 $\boldsymbol{s}(t)$ を観測し，アクターネットワークにより，行動 $\boldsymbol{a}(t)$ を決定する．

(3) エージェントが行動 $\boldsymbol{a}(t)$ を実行することにより，状態が $\boldsymbol{s}(t+1)$ に遷移する．

(4) クリティックは環境の状態 $\boldsymbol{s}(t+1)$ から報酬 $r(\boldsymbol{s}(t+1))$ を受け取り，アクターへTD誤差 δ を出力する．

$$\delta = r(\boldsymbol{s}(t+1)) + \gamma V(\boldsymbol{s}(t+1)) - V(\boldsymbol{s}(t)) \quad (1)$$

Reinforcement Learning by KFM Associative Memory based on Area Representation with Refractoriness
Atsushi Shimizu and Yuko Osana (Tokyo University of Technology, osana@cc.teu.ac.jp)

ここで，γ $(0 \leq \gamma \leq 1)$ は割引率，$V(s(t))$ は状態 $s(t)$ に対する状態価値関数である．

(5) 適格度 $e_t(s)$ を用いてすべての状態 $s(s \in S)$ について，状態価値 $V(s)$ を次のように更新する．

$$V(s) \leftarrow V(s) + \xi \delta e_t(s) \qquad (2)$$

ここで，ξ $(0 \leq \xi \leq 1)$ は学習率である．

(6) 適格度 $e_t(s)$ を以下のように更新する．

$$e(s) \leftarrow \begin{cases} \gamma\lambda e(s) + 1 & (s = s(\mathrm{t+1}) \text{ のとき}) \\ \gamma\lambda e(s) & (s \neq s(\mathrm{t+1}) \text{ のとき}) \end{cases} \qquad (3)$$

ここで，γ $(0 \leq \gamma \leq 1)$ は割引率，λ はトレース減衰パラメータである．

(7) TD 誤差に基づいてアクターネットワークの重みを更新する．ここでは (2) で観測した状態 $s(t)$ とそれに対してとった行動 $\boldsymbol{a}(t)$ より学習ベクトル $\boldsymbol{X}^{(t)}$ を生成し，学習を行う．

(a) TD 誤差が 0 より小さいとき

状態 $s(t)$ に対して行動 $\boldsymbol{a}(t)$ をとったときにそれが望ましくないと判断されたときには，勝ちニューロンの重みをランダムに初期化する．勝ちニューロンの重みが固定されている場合には固定を解除する．

(b) TD 誤差が 0 より大きいとき

状態 $s(t)$ に対して行動 $\boldsymbol{a}(t)$ をとったときにそれが望ましい行動であると判断された場合には，$\boldsymbol{X}^{(t)}$ を用いて学習を行う．

$$\Delta \boldsymbol{W}_i(t) = \delta H(d_i) h_{ri} (\boldsymbol{X}^{(t)} - \boldsymbol{W}_i(t)) \qquad (4)$$

ここで，$H(d_i)$ は準固定を行うための係数，$\alpha(t)$ は学習係数，h_{ri} は近傍関数である．また，$d(\boldsymbol{X}^{(t)}, \boldsymbol{W}_r) \leq \theta^l$ が満たされたとき，勝ちニューロン r に結合する重み $\boldsymbol{W}_r$ を固定する．

(c) TD 誤差が 0 のとき

TD 誤差が 0 のときは重みの更新は行わない．

(8) (2) に戻る．

4 計算機実験

提案手法の動作と有効性を確認するために，経路選択問題を題材として実験を行った．

実験では，有限で離散的な 2 次元空間にエージェントを配置し，適切な経路を選択することができるかを検証した．2 次元空間にはゴールが存在し，エージェントはゴールにたどり着くことを目的として行動する．エージェントはエージェント自身が存在する位置からの環境を状態として観測することができ，前進，後退，左回り，右回りの 4 パターンの行動をとるものとする．

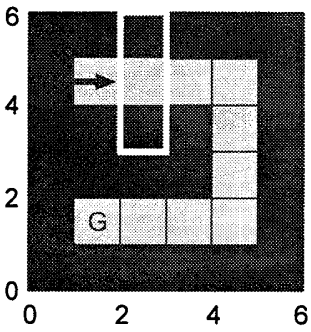

図 1: 環境とエージェント

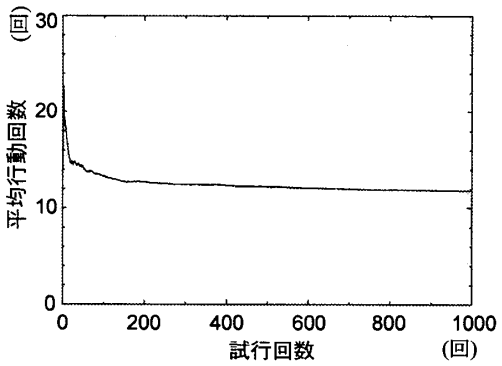

図 2: ゴールまでの平均行動回数

図 1 のような環境で実験を行った．この図において，矢印がエージェント，エージェントの前の白い枠がエージェントの観測可能範囲を表している．エージェントの初期位置を (1, 4) とし，エージェントは (1, 1) のゴールにたどりつくことを目的として行動する．

図 2 にエージェントがゴールにたどり着くまでの平均行動回数の変化を示す．この結果より，適切な行動へと学習が収束することが確認できた．

参考文献

[1] R. S. Sutton and A. G. Barto : Reinforcement Learning, An Introduction, The MIT Press, 1998.

[2] I. H. Witten : "An adaptive optimal controller for discrete-time Markov environments," Information and Control, Vol.34, pp. 286–295, 1977.

[3] T. Imabayashi and Y. Osana: "Implementation of association of one-to-many associations and analog patterns in Kohonen feature map associative memory with area representation," Proceedings of IASTED AIA, Innsbruck, 2008.