2004－MPS－49　(6)
2004／5／7

エピソード記憶による迷路課題の学習アルゴリズムの提案

青田 佳士[1,2,3]　山口 陽子[2,3]

TD learning といったナビゲーションによく利用される学習モデルでは一般に，Alternation maze task のようにゴールの位置が交互に変わるような迷路課題において，学習の干渉をいかに克服するかが問題となっている．これまで提案されたモデルの多くは，個々の状況を分離し学習の干渉を避ける形で課題に取り組むができても，さらにゴールの移り変わり方そのものを学ぶといった過去の経験を活用するような学習は困難であった．本研究では，モデルが経験・蓄積したエピソード記憶を状況に応じて想起し実行するのみでなく，エピソードの組み合わせも探索することでゴールの位置の移り変わりのルールも学習する新しいアルゴリズムを提案する．

A learning algorithm for solving maze tasks with episodic memory

Yoshito Aota[1,2,3]　and　Yoko Yamaguchi[2,3]

It is known that how to conquer the interference of learning is one of big problem when we use normal learning theory of navigation like TD learning method for solving the alternation maze task. Alternation maze task is the task whose goal position changes alternatively. Although many models could avoid meeting the interference of learning by separation of each situation, they could not learn the rule of goal transition itself. They showed difficulty in utilization of past experience to the next situation. In this research, we propose new learning algorithm for distinguishing correct situation with episodic memory. Moreover, we show this algorithm can learn correct order of goal transition by searching better combination of episodes.

1. はじめに

TD learning といったナビゲーションによく利用される学習モデルでは一般に，Alternation maze task [1]のようにゴールの位置が交互に変わるような迷路課題において, 学習の干渉(「カタストロフィック干渉」[2]）をいかに克服するかが問題となっている．Alternation maze task では通常，ゴールの位置の変化を除いて環境の変化はないため，モデルは同じ入力に対して異なる出力を適切に実行できなければ学習の干渉は避けられない．

追加学習[3],[4]や行動履歴の木構造によるアプローチ[5]などこれまで提案されたモデルの多くは，個々の状況を分離し学習の干渉を避ける形で課題に取り組むができても，さらにゴールの移り変わり方そのものを学ぶといった過去の経験を状況依存的に活用するような学習は困難であった．

本研究では，モデルが経験・蓄積したエピソード記憶を状況に応じて想起し行動するのみでなく，エピソード記憶の組み合わせも探索することでゴールの位置の移り変わりのルールも学習する新しいアルゴリズムを提案する．

1 横浜国立大学
Yokohama National University
2 独立行政法人 科学技術振興機構（JST）
Japan Science and Technology Agency (JST)
3 理化学研究所
RIKEN

2. 学習アルゴリズム

2.1. モデルの基本方略と実際のルール

本節では，エピソード記憶を利用して迷路課題を解くためのモデルの基本方略を述べる．

ここで一つのエピソード記憶は，成功した時，つまりゴールに着いて報酬を得た時点，もしくは失敗した時，つまりゴールを予測してはずれた時点を区切りとしてその間に経験した入力と出力の一連の系列（これを行動と呼ぶことにする）や得られた報

酬のタイミングの記憶とする．以上はスタートからゴールまで（もしくは失敗まで）の 1 試行分の記憶であるが，後述するように複数試行分を一つのエピソードとして統合して取り扱うこともある．

本モデルでは，まず大まかに述べると以下のような流れで学習が進む．

1. エピソードの選択
   状況（前回のエピソード，外部入力）に応じて貯蔵しているエピソードを想起し，より報酬の得られる度合いの高いエピソードを選択
2. エピソードの実行
   選択されたエピソードの成功部分(報酬を得た行動)を実行，失敗部分は探索行動に切り替え
3. 新規エピソードの作成
   新奇行動は新規エピソードとして貯蔵
4. エピソード間の順序関係の学習
   前回のエピソードから今回のエピソードへの結合を強化
5. エピソードの統合
   報酬の取得成功が続いた場合,それらを一つのエピソードとして統合

このような基本方略を実現するために設けた実際のルールは下記のとおりである．

I. スタート前にエピソードを選択
   a 前回のエピソードと正に結合したエピソードがある場合，それらのうち内的評価値（報酬の数，短い経路）の高いエピソードを選択
   b 前回のエピソードと正に結合したエピソードがない場合，前回のエピソードと負に結合したエピソード以外で，外部入力（モデルの現在位置）と結合のあるエピソードで内的評価値の高いものを選択
   c 以上のいずれにも該当するエピソードがない場合，探索行動を実行
   d エピソードの選択過程にランダムネスを加え，探索行動も適度に起こるようにする
II. エピソードの実行と新規エピソードの作成と統合、およびエピソード間の結合
   a 選択エピソードのスタートから最後に報酬を得た時点までの行動と同じ行動を実行する
   b 選択エピソードどおりに全ての報酬を得た場合，c へ
   c 前回のエピソードから選択エピソードへの結合強度を正にして試行終了
   d 探索行動の結果報酬を得た場合，もしくは選択エピソードとは別の時点で報酬を得た場合，その行動と同じエピソードがあればそれを，なければ新規にエピソードを作成してそれを選択エピソードとし c へ．前回のエピソードから選択エピソードへの結合は負にする．
   e 選択エピソード作成時には得られた報酬を今回は得られなかった(失敗した)場合，前回のエピソードから選択エピソードへの結合強度を負にする．今回の行動と同じエピソードがある場合はそれを，ない場合は新規にエピソードを作成してそれを選択エピソードとし，前回のエピソードから選択エピソードへの結合強度を 0 にする．前回の失敗と（今回が初めての失敗の場合は、1 回目の試行と）今回の失敗の間に 2 回以上報酬を得た行動があり，それと同じエピソードがある場合はそれと，なければその行動をコードするエピソードを新規に作成しそれと，前回の失敗エピソードとの結合を正にする．I へ

2.2. 迷路課題

本研究では，モデルの計算実験に用いる迷路課題として T 字型と十字型の 2 種類のプラットフォームを用意し（図 1），以下の 1 から 3 の 3 つの課題をモデルに行わせた．いずれの課題も環境の変化としてはゴールの位置の変化のみを想定し，モデルの行動に応じてゴールの位置が変わる．モデルは

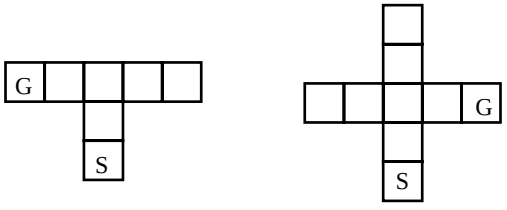

図 1. 迷路課題のプラットフォーム

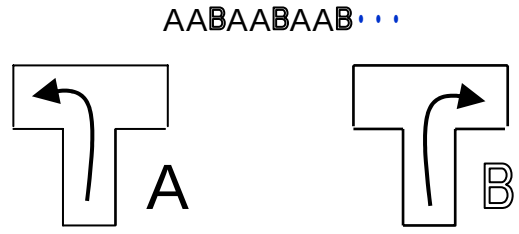

図 2. T 字型で 1 周期 3 試行の迷路課題

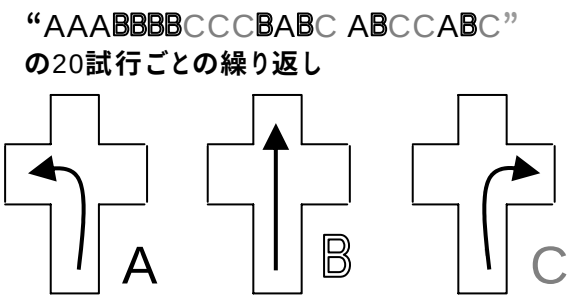

図 3. 十字型で 1 周期 20 試行の迷路課題

スタート地点からゴールまでの道のりのみでなく，ゴールの位置の変化則をも獲得しなければ適切にゴールの位置を予測することはできない．また，外部入力としてプラットフォーム上のモデルの位置を与えた．

図 1 のプラットフォームにおいて，S はスタート地点，G はゴール地点を示す．G の位置はモデルの行動に応じて S 以外の道端に移動する．各試行において、モデルは S から出発して道端にたどり着くたびに S 地点に戻される．G 地点にたどり着いた時点で 1 試行終了とし，報酬が与えられる．1 マスの移動を 1 ステップとし，ゴールにたどり着くまで試行は続く．そのため 1 試行あたり 4 ステップ以上を要することになる．また後方への移動はないとし，モデルは前方か左右の 3 方向への移動しかできないものとする．

＜実験条件 1＞

T 字型迷路．ゴールの位置はモデルがゴールにたどり着くたびに図 2 のように A→A→B→A→A→B→...と変わる．つまり 3 試行を 1 周期としてゴールの位置が変わる．モデルは，A の行動をとって報酬を得た次の試行で再び A の行動を実行し，次に B の行動をとって報酬を得るといった行動系列を学習しなければならない．

＜実験条件 2＞

T 字型迷路．実験条件 1 の課題を行う前に，ゴールの位置が A に固定されている場合を 200 試行学習後，B に固定されている場合を 200 試行学習する．

＜実験条件 3＞

十字型迷路．ゴールの位置はモデルがゴールにたどり着くたびに図 3 のように 20 試行を 1 つの周期として順に変わる．

＜実験条件 4＞

十字型迷路．実験条件 3 の課題を行う前に，ゴールの位置が A に固定されている場合を 200 試行学習後，B に固定されている場合を 200 試行，さらに C に固定されている場合を 200 試行学習する．

3. シミュレーション結果

T 字型迷路の結果を図 4，5 に示す．図 4 は学習曲線で，1000 個体分のモデルの各周期の平均失敗数を縦軸にとり，横軸に周期数を示した．黒線が実験条件 1 で，グレーの線が実験条件 2 である．この課題を適切に学習するには，図 2 で同じ A の行動系列を内的に区別する必要があり、モデルはこれを自律的に獲得しなければならない．図 4 より，実験条件 1，2 とも 30 周期程度で誤ったエピソードを選択しなくなり，内的区別をうまく実現できていることが分かる．図 5 は T 字型迷路課題の各周期の各試行において，モデルがゴールに到るまでに要したステップ数の 1000 個体分の平均である．黒線とグレーの線はそれぞれ実験条件 1 と 2 を示し，点線はランダムな行動をとり続けた場合の結果を示す．実験条件 2 では部分的な状況を含むとはいえ別の状況を事前に経験した後でも学習の干渉を起こさず，むしろ実験条件 1 とくらべて事前の経験が活かされていることが分かる．しかしながら，両条件とも最短ステップである 4 には到らず，局所解で収束している．

図 6，7 は十字型迷路の結果で，図 4，5 と同様に学習曲線とゴールまでの平均ステ

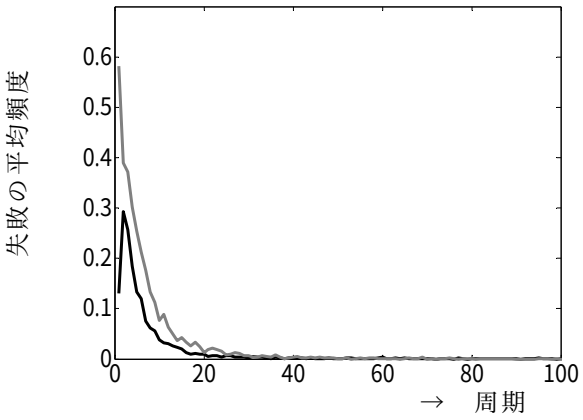

図 4. T 字型迷路の各周期における失敗の頻度

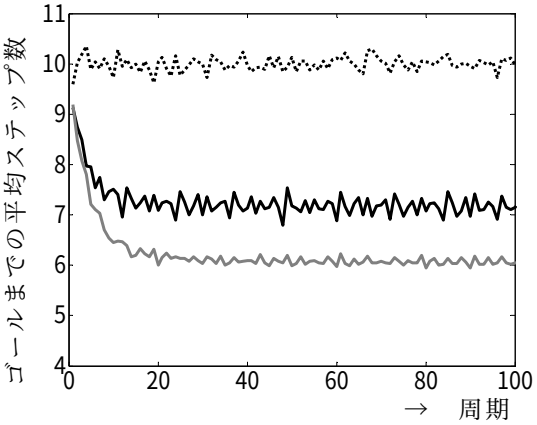

図 5. T 字型迷路の各周期におけるゴールまでの平均ステップ数

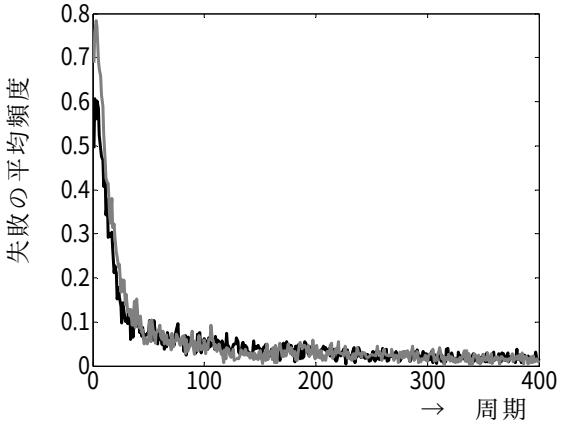

図 6. 十字型迷路の各周期における失敗の頻度

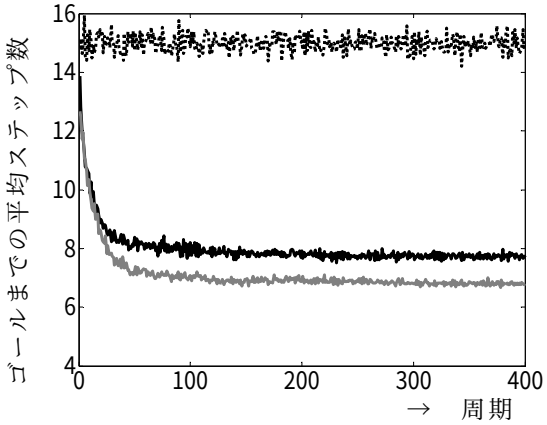

図 7. 十字型迷路の各周期におけるゴールまでの平均ステップ数

ップ数を示した．図 6, 7 において黒線とグレーの線はそれぞれ実験条件 3 と 4 を表し，図 7 において点線はランダムな行動をとり続けた場合の結果を示す．

図 6，7 とも図 4，5 と同様の結果が得られた．そのため，1 周期が 20 試行と非常に長い系列なだけでなく分岐点におけるモデルの選択肢が 3 つに増えた難しい課題でも，モデルは十分に学習の収束能力を示し，かつ別の状況の学習経験を活かすことも可能であることが分かる．

4. まとめ

エピソード記憶を利用した新しい学習アルゴリズムを提案した．Alternation maze task に適用した結果，モデルは複雑な課題にも適切なエピソードを選択でき，過去に学習した別の状況もより適切な解の探索に役立っていることが分かった．依然，局所解に収束しているため，今後は過去の経験が新奇な状況にどう影響するかを詳細に調べたい．

文　献

[1] E.R. Wood, P.A. Dudchenko, R.J. Robitsek, and H. Eichenbaum, Neuron, Vol. 27, pp.623-633, 2000.

[2] 都築誉史，河原哲雄，楠見　孝，心理学研究，72 巻，6 号，pp. 541-555, 2002.

[3] 小林正宣，小澤誠一，阿部重夫，計測自動制御学会論文集，Vol. 38, No. 9, pp.792-799, 2002.

[4] 大平岳将，山内康一郎，大森隆司，信学技報，NC2001-178, pp.79-85, 3 月，2002.

[5] R.A. McCallum, Proc. 12th International Conf. on Machine Learning, pp.387-395, 1995.