F-031

環境オブザーバと方策オブザーバを用いた変動環境にロバストな学習モデル

A robust learning model to environmental change with an environmental observer and a policy observer

小野　将寛†　　**佐々木　守**　　**岩田　穆**
Masahiro Ono　Mamoru Sasaki　Atsushi Iwata

１．はじめに

未知なる変動環境に対して，ロボットが学習するための手法として，複数のエキスパートモジュールを環境に応じて切り替えるモジュール構造を持つ学習モデルがJacob らのモデル[1]をはじめとして複数提案されている．

その中の代表的なものに，銅谷らによって提案されたMMRL(Multiple Model-Based Reinforcement Learning) [2]がある．このモデルでは，各モジュールは環境の予測モデルと行動を学習する強化学習コントローラにより構成される．そして，予測モデルの正確性に基づいた責任信号によって各モジュールの出力への貢献度が決まる．また，予測モデルや価値関数の更新にも責任信号が使用される．

しかし，環境間に高い類似性がある場合には，複数のモジュールの責任信号が増加し，それらが競合することにより，行動価値関数が理想値から低下する場合がある．また，モジュールの追加機能がないので，環境数が増加した場合の環境への適応が困難となる．

以上を踏まえ，本研究ではモジュラー学習を基本とした単一モジュール選択構造のモデルを提案する．また，この構造を採用することにより，ハードウェア化を考えた場合に，同時に複数のモジュールが活性化することがないので消費電力の低減も期待できる．

本モデルでシミュレーション実験を行い，その有効性を示す．

２．提案モデル

図 1 に提案モデルの概略を示す．モジュールは大きく分けて方策 $f_m(s)$と環境の予測モデル $P_m(s', s, a)$から構成される（m: モジュール番号, s': 次時刻の状態，s，a: 現在の状態と行動）．f_m は現在の状態に対して適切な行動を出力する関数である．P_m はモジュール m が担当する環境の状態遷移確率を表す．そして，これらは方策オブザーバと環境オブザーバの出力に基づいて更新される（2. 5 参照）．ただし，状態遷移確率は 0，1 の 2 値とする．

図 2 に示すように，モデルの動作は，a. 環境変動の検出(2. 3 参照)，b. 現在の環境に対して有効なモジュールの選択(図 1，2. 4 参照,)，c. モジュールの追加/統合(2. 5 参照)から構成される．なお，紙面の都合上 c. の概略図は省略した．

2. 1　環境オブザーバ

環境オブザーバは現在の環境の状態遷移確率を観測し $P_{obs}(s', s, a)$として記憶する．ただし，環境変動の検出後は新しい環境の状態遷移確率を観測するためにリセットされる．

†広島大学大学院先端物質科学研究科，ADSM

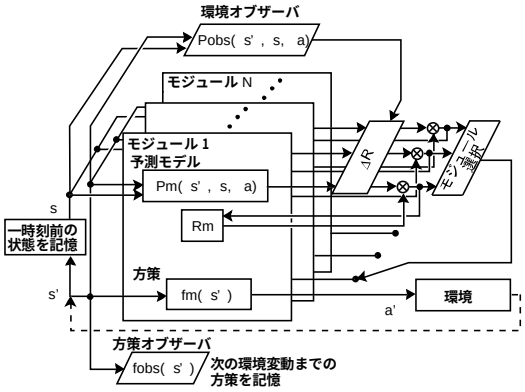


図 1: 提案モデル(モジュール選択時の動作)

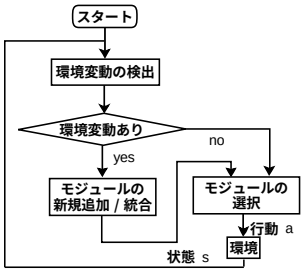


図 2: 提案モデルの動作

2. 2　方策オブザーバ

方策オブザーバは現在の環境に有効な方策を $f_{obs}(s)$に記憶する．詳しくは，無効ルール（一時刻前と同じ状態に遷移する行動）を経験せずにゴールに到達できた場合のみ，その方策を記憶する．ただし，環境オブザーバと同様の理由で環境変動検出後はリセットされる．

2．3　環境変動の検出

環境が変動した場合は P_{obs}が 0→1 または，1→0 へ変動するので，その時点で環境変動を検出する．

2．4　モジュールの選択

現在の環境に適したモジュールを選択するために，選択評価変数 R_mを導入する．R_mは 0, 1 の 2 値であり，R_m=1 となるモジュールが選択される（複数存在する場合にはその中からランダムに選択する．）．

R_m の計算を説明する．環境が状態遷移する度に P_{obs} と P_m を比較し，一致すればΔR_m =1，異なればΔR_m =0 とし，$R_m \leftarrow R_m \times \Delta R_m$ を計算する．ただし，環境変動を検出後は，R_m を全てリセット(1 に)する．（図 1 参照）

2．5　モジュールの追加・統合

環境変動を検出後，環境オブザーバによって観測された P_{obs} と方策オブザーバによって観測された方策 f_{obs} を組

としたモジュールを新規に追加するか，あるいは既存のモジュールへ統合するかを判断する．その条件は以下とする．

・追加: 全てのモジュールに対して $P_{obs} \neq P_m$ となる $P_m(s', s, a)$が少なくとも一つ存在する．

・統合: $(P_{obs} = P_m)$ かつ $(f_{obs} \neq f_m$以外)

統合する際は，P_{obs} と P_m，f_{obs}と f_mの和集合をとる．また，統合条件で f_{obs} を考慮するのは，統合するモジュールの選定精度を向上させるためである．

2．5　行動

選択されたモジュールの方策に従う．ただし，方策が決定されていない状態の場合はランダムに行動する．

3．シミュレーション実験

3．1　問題設定

表 1 の類似度を持つ 6 種類の 5x5 サイズの迷路(図 3)を用意し，周期 T(>25)で変化させる．エージェントは上下左右に行動できる．ゴールに到達すると試行が終了し，スタートに戻されて試行が開始される．類似度は，迷路間で最適方策が一致する個数を全方策数で正規化することにより求めた．初期状態ではモジュール数は 0 とする．

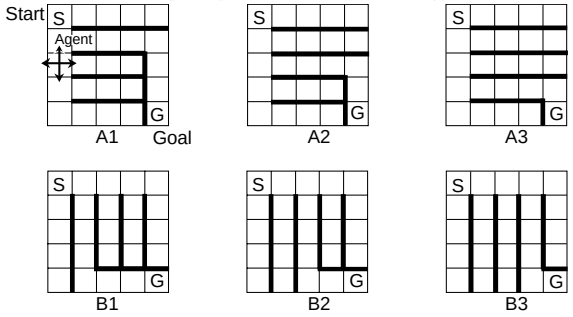


図 3: 迷路の種類

表 1: 各迷路間の類似度

(%)

	A1	A2	A3
B1	25	21	17
B2	21	17	13
B3	17	13	8

(a) **異種間**

(%)

	A1(B1)	A2(B2)	A3(B3)
A1(B1)		64	56
A2(B2)	64		64
A3(B3)	56	64	

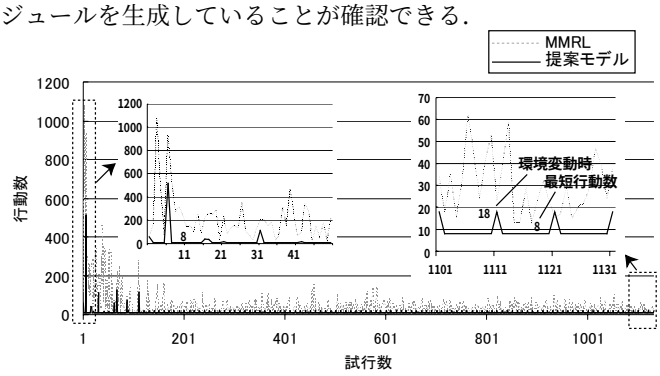
(b) **同種間**

3．2　実験結果

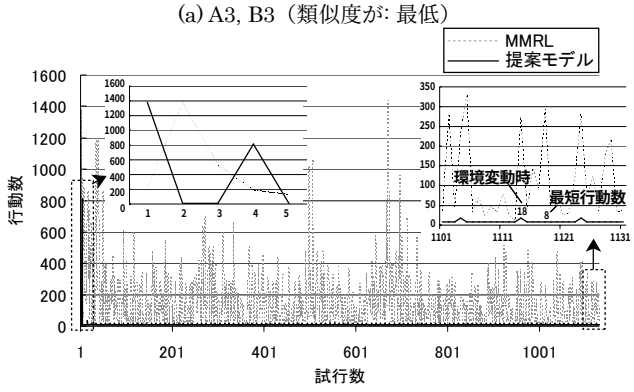
環境の類似性に対しての評価を図 4 に，環境変動の周期 T に対しての評価を図 5 示す．比較対象として MMRL の結果も載せた．MMRL の行動選択には[2]と同様に Gibbs 分布によるソフトマックス手法を使用した．忘却係数 α: 0.9，逆温度パラメータ β:1，割引率 γ: 0.9，ゴールした場合の報酬は 10 とした．

図 4 (a), (b)では，類似度の低い A3, B3(8%)と類似度の高い A1, A2(64%)の 2 迷路を 100 回行動ごと(周期 T=100)に切り替えた時に，スタートからゴールまでの一試行中に要した行動数の変移を示している．MMRL が類似度に依存して行動数が大きく変化しているのに対し，提案モデルは類似度に依存せずどちらも 200 試行以内に最短行動数に落ち着いている．

図 5 は，6 迷路を周期 T で変化させた場合に，10^5 回行動するまでに獲得した報酬を示す．報酬量の比から，本モデルが MMRL よりも約 5 倍～8 倍速くエキスパートモジュールを生成していることが確認できる．



(a) A3, B3（類似度が: 最低）



(b)　A1, B1（類似度: 最大）

図 4: 2 迷路を 1000 行動ごとに切り替えた場合の結果．（ (a): A3, B3（類似度: 最低）, (b): A1, B1（類似度: 最大） ）

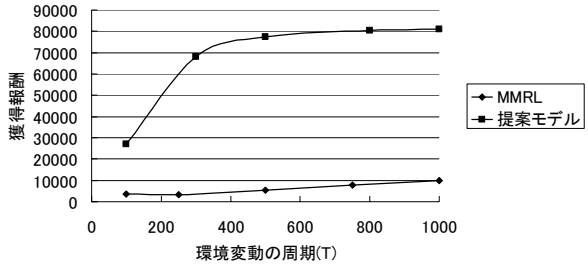


図 5: 獲得報酬 （一定時間内*で 6 迷路を周期 T で変化させた場合，*: エージェントが 10^5 回行動）

4．むすび

環境変動にロバストな学習を目指して，環境オブザーバと方策オブザーバを導入したモジュラー学習を提案した．そして，迷路問題を例に，環境間の類似性，環境変動の周期を変えて実験を行い，モデルの有効性を示した．

参考文献

[1]: R.A. Jacobs, M.I. Jordan, S.J. Nowlan, and G.E. Hinton, “Adaptive mixtures of experts”, Neural Computation, vol.3, pp.79-87, 1991.

[2]: Kenji Doya Kazuyuki Samejima Ken-ichi Katagiri and Mitsuo Kawato, “Multiple Model-Based Reinforcement Learning”, Neural Computation, vol.14, pp.1347-1369, 2002.