Vol.2014-MPS-97 No.16
2014/3/4

リンケージ型学習分類子システムによる類似環境に適応可能な汎用的知識の獲得と活用法の提案および性能評価

臼居 浩太郎[1,a)]　中田 雅也[2,b)]　髙玉 圭樹[2,c)]

概要：本研究では，類似する環境に適応可能な遺伝的機械学習手法としてリンケージ型学習分類子システム (XCSAM with Linkage-Classifier: XCSL) を提案する．具体的には，行動最適性の高い分類子同士の結びつきを表現したリンケージ型分類子 (Linkage-Classifier) を形成し，行動選択に利用する．提案手法の有効性を検証するために，Multi-step 問題の一般的なベンチマーク問題である Block World 問題を用いて，従来手法 XCS, XCSAM との比較をしたところ，類似する環境に変化する場合において，リンケージ型分類子を活用することで XCS に対しては約 18%，XCSAM に対しては約 25%の学習回数で新しい環境に適応可能であることが明らかとなった．

1. はじめに

近年の強化学習 [10][11] の発展に伴って，多自由度ロボット等が扱う高次元状態行動空間において，適切な行動制御則の自律的獲得を課題とする研究が盛んに行われている [8][9][6]．この課題の解決として，学習効率化の観点より，複数状態に適用可能な汎用的行動制御則を獲得する方法が考えられる．この方法は，適切な行動制御則を獲得するまでに多くの学習回数を要することが懸念されるが，次の 2 つの重要な利点が挙げられる．まず，1) 環境が変化した場合でも，変化後の環境が類似する状態行動空間である場合，学習した行動制御則が再適用可能である．次に，2) 汎用的行動制御則は，状態行動空間に存在する特徴的な情報 (知識) を抽出したものであることから，状態行動空間の理解に寄与する．このような利点から，ロボット制御問題等を扱う逐次的意思決定問題 (Sequential Decision Task) における汎用的行動規則の獲得手法の構築が重要な課題である．

本研究では，汎用的行動規則の獲得手法として学習分類子システム (Learning Classifier System: LCS)[4] を用いる．LCS は強化学習と遺伝的アルゴリズム (Genetic Algorithm: GA)[3] から構成され，行動制御則に対応する分類子 (if-then ルール) を一般化することで，汎用的行動制御則を学習可能である．特に，現在主流である正確性に基づく学習分類子システム (Accuracy-based LCS : XCS)[12] は，得られる報酬を正しく予測した汎用的行動制御則が獲得可能である [12]．しかし，XCS は全状態行動空間を網羅的に学習するため，XCS が膨大な試行回数を要する問題 [1] が存在する．この問題の解決に向けて，目的達成に必要な行動制御則 (最適行動) のみ学習する LCS (XCS with Adaptive Action Mapping: XCSAM)[7] は，XCS よりも少ない試行数で適切な行動制御則を学習可能であることが示されている [7]．しかしながら，XCS や XCSAM は正確性の指標を用いることで，環境変化時に学習した汎用的行動制御則が再適用できないという限界が存在する．

上記の問題を解決するために，我々は，類似する環境に変化した場合では，行動制御則の行動最適性は維持される傾向があることに着目し，最適行動である行動制御則の実行順を抽出するリンケージ型学習分類子システム (XCSAM with Linkage-Classifier: XCSL) を提案する．

2. 行動最適性に基づく LCS：XCSAM

XCSAM で用いる分類子は，条件部 (if) ならびに行動部 (then) と，予測報酬 p や適合度 F，経験値 exp，行動最適性 eam などの評価値により構成される．XCSAM は環境から知覚した状態に照合する分類子を分類子集団 (population:$[P]$) より選択し，照合集合 (Match set:$[M]$) を形成する．そして各行動に対する行動価値を式 (1) により計算し予測配列 (prediction array) を形成する．ここで $P(s_t, a)$ は状態 s_t における行動 a の行動価値である．

[1] 電気通信大学 情報理工学部
The University of Electro-Communications, Tokyo, Japan
[2] 電気通信大学大学院 情報理工学研究科
Graduate School of Informatics and Engineering, The University of Electro-Communications, Tokyo, Japan
[a)] usui@cas.hc.uec.ac.jp
[b)] m.nakata@cas.hc.uec.ac.jp
[c)] keiki@inf.uec.ac.jp

ⓒ 2014 Information Processing Society of Japan

$$P(s_t, a) = \sum_{cl_k \in [M](a)} p_k \times \frac{F_k}{\sum_{cl_i \in [M](a)} F_i} \quad (1)$$

この行動価値に比例して行動選択を行った後，選択した行動を持つ分類子を $[M]$ から選択し，行動集合 (action set:$[A]$) を形成する．その後，行動を実行し報酬 r を獲得する．次に $[A]$ 内の分類子について各評価値のパラメータを更新する．特に予測報酬 p は下記の式 (2) に示すように，獲得報酬 r と最大行動価値 $maxP(s_t, a)$ を用いて計算する．ここで，パラメータ $\beta(0 \leq \beta \leq 1)$，$\gamma(0 \leq \gamma \leq 1)$ はそれぞれ学習率，割引率と呼ばれ，学習の更新速度と将来の報酬を考慮する度合いを制御する．

$$p_j \leftarrow p_j + \beta(r + \gamma maxP(s_t, a) - p_j) \quad (2)$$

次に，実行した行動について最適行動の識別を行う．状態 s_t における最適行動は最大行動価値 $maxP(s_t, a)$ を持つ行動 a であり，状態遷移時に目的達成 (報酬獲得) に近づくほど $maxP(s_t, a)$ は増加する．最適行動実行時に移動した状態 s_{t+1} における $maxP(s_{t+1}, a)$ は，$maxP(s_t, a)$ より $1/\gamma$ だけ増加する．よって，$maxP(s_{t+1}, a)$ が $\zeta \times maxP(s_t, a)/\gamma$ 以上であった場合，実行した行動が最適行動であることが識別できる．ここで，ζ は誤差許容率であり，$maxP(s_{t+1}, a)$ の収束誤差を許容する度合いを制御する．次に，上記の条件を用いて分類子の行動部に最適行動をもつ分類子 (最適行動ルール) を特定し，削除する分類子を決定する．このために，eam を用いて分類子の行動部が最適行動であるかを判定する．具体的には eam を式 (3) より更新する．ここで nma は環境における取りうる全行動数を表す．式 (3) より eam が 1 に収束するとき，その分類子は最適行動ルールとなり，一方で nma に収束するとき，非最適行動ルールと識別する．

$$eam \leftarrow \begin{cases} eam_i + \beta(1 - eam_i) \\ \quad if\ maxP(s_{t+1}, a) \geq \zeta \times maxP(s_t, a)/\gamma \\ eam_i + \beta(nma - eam_i) \quad otherwise. \end{cases} \quad (3)$$

最後に，遺伝的アルゴリズムを用いて分類子の生成と削除を行い $[P]$ を進化させる．その際，行動の最適性を表す評価値 eam に応じて最適行動でないルールを優先的に削除し，生成を抑制することで，最適行動のみを学習する．

3. リンケージ型学習分類子システム：XCSL

図 2 に XCSL のアーキテクチャの概要を示す．XCSL は先行研究である XCSAM に新たに 1) リンケージ型分類子 (Linkage-Classifier) の形成法および解除法，2) リンケージ型分類子を考慮した行動選択法を導入する．

3.1 リンケージ型分類子の形成

リンケージ型分類子を形成する，最適でかつ同一な行動を持つ分類子を格納するために，最適行動集合 $[BA]$ を用

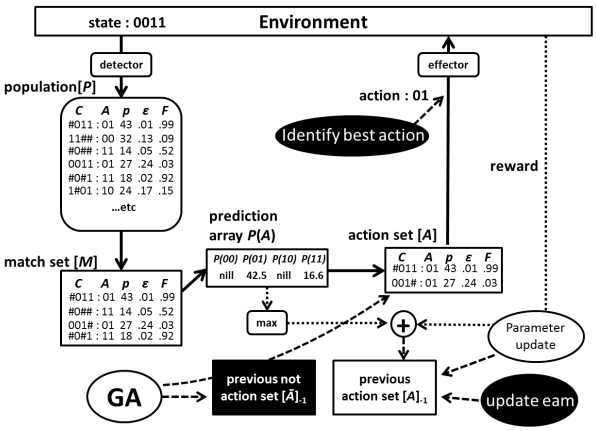


図 1 XCSAM のアーキテクチャの概要

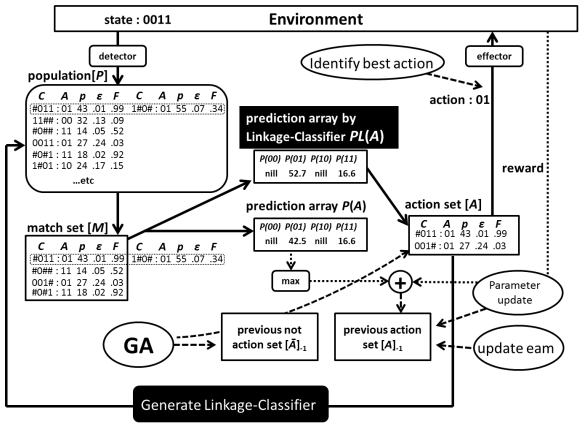


図 2 XCSL のアーキテクチャの概要

いる．$[BA]$ に行動集合 $[A]$ を格納する条件は 1) 状態 s_t の行動 a が最適行動であり，かつ，2) $[BA]$ 内の行動と行動 a が一致する場合である．上記の条件が満たされなくなった時点で，$[BA]$ に格納されている各行動集合 $[A]$ 内の分類子からリンケージ型分類子の形成を行う．ここで，行動集合が 1 つである場合は，リンケージ型分類子の形成は行わない．図 3 に，リンケージ型分類子を構成する分類子を，$[BA]$ から選択するメカニズムを示す．$[BA]$ に格納されている各行動集合 $[A]_t$ において，各分類子の評価値がそれぞれ $exp > \theta_{exp}$, $error < \epsilon_0$, $eam < \theta_{eam}$ を満たす分類子の中から，最小の eam を持つ分類子を選出する．その後，連続する状態の分類子を用いてリンケージ型分類子を形成する．ただし，連続する分類子がない場合，リンケージ型分類子を形成しない．ここで，θ_{exp} は，分類子が生成されたばかりでないことを判断するパラメータであり，ϵ_0 は，従来手法の XCS 及び XCSAM において分類子の適合度 Fitness の計算を行うのに使用するパラメータである．θ_{eam} は，分類子の行動が最適であるかを判断するパラメータである．θ_{exp} と θ_{eam} は提案手法から新たに設定されるパラメータであり、以上のパラメータは，問題に応じて適切に設定される．最後に，リンケージする条件の一つであ

る $eam < \theta_{eam}$ を満たさなくなったリンケージ型分類子はリンケージを解除し，ただの分類子とする．このとき，[P] で淘汰された分類子については，リンケージを解除せずリンケージ型分類子として保持し続ける．

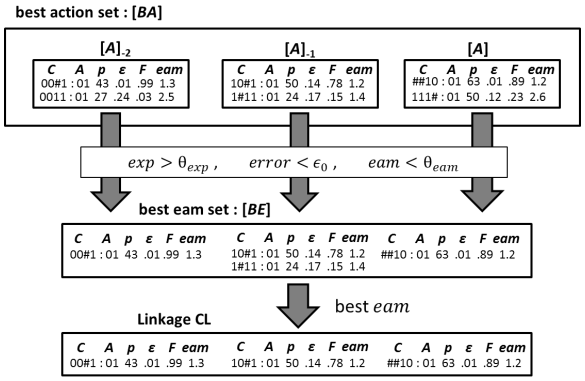


図 3 リンケージ型分類子の形成法

3.2 リンケージ型分類子による行動選択

行動決定のみに使用する予測配列 (prediction array by Linkage-Classifier: $PL(s_t, a)$) を新たに導入する．$PL(s_t, a)$ では，各予測報酬値の計算を行う際にリンケージ型分類子を考慮して計算を行う．具体的には，従来手法では [M] 内の分類子の評価値から予測値を算出していたのに対して，提案手法では，[M] 内の分類子がリンケージ型分類子を構成する分類子の一つである場合，そのリンケージ型分類子の分類子の中で，最後にリンケージされた分類子のパラメータを用いて計算を行う (図 4)．評価値の更新を行う予測配列 (prediction array: $P(s_t, a)$) は，XCS, XCSAM 同様に持っており，同様に算出を行う．

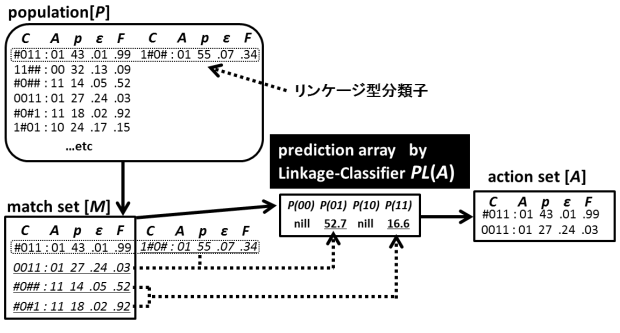


図 4 リンケージ型分類子による行動選択

4. 評価問題

提案手法である XCSL の学習性能を評価するために，Multi-step 問題の一般的なベンチマーク問題である Block World 問題 [12] を用いる．

4.1 Block World 問題

Block World 問題は難易度の異なる迷路において，*animat* と呼ばれるエージェントが各セル間を遷移し餌 (報酬) の獲得を目的とする問題である．迷路は通路（Empty position : " "），障害物（Obstacle : "T"），餌（Food : "F"）で構成され，エージェントは現在地を中心に 8 近傍を知覚し 8 方向に移動可能である．各セルは障害物が 01，通路が 00，餌が 11 で表され，状態を表すコードは現在位置を中心として真上のセルから時計回りにコーディングすることで，長さ 16 のビットで与えられる．障害物方向へ移動する場合は，移動不可としてその状態に留まる．餌に到達した場合は報酬 $r = 1000$ を得て探査を終了する．ただし，探索ステップ数が最大ステップ数を超えた場合も探索を終了する．スタート位置はランダムである．本研究で使用する迷路問題は，Maze5[5], Maze6[5], Maze6-a, Maze6-c である．

5. 実験

5.1 実験内容

XCSL の環境変化時の再学習性能を，類似した環境への環境変化に関して評価を行う．実験では，表 1 および図 5 の 4 通りの環境変化を想定して実験を行う．図 5 の赤丸は環境の変化した部分である．環境変化 I および環境変化 II は，餌までの経路が増減したときの学習性能を検証する．環境変化 III は，環境変化 I と同様に餌までの経路が減る環境変化であるが、環境変化の起こる位置による影響の違いを検証する．環境変化 IV は，リンケージ型分類子の獲得状況による学習性能への影響を検証する．

変化前の環境に対して 10000 回の学習を行った後，環境を変化させ，さらに 10000 回の学習を行う．XCSL の性能を検証するために，従来手法 XCS, XCSAM と性能を比較する．

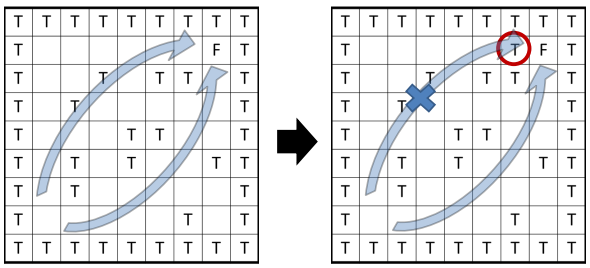
表 1 類似環境への環境変化の種類

I)	Maze6-a	→	Maze6
II)	Maze6	→	Maze6-a
III)	Maze5	→	Maze6

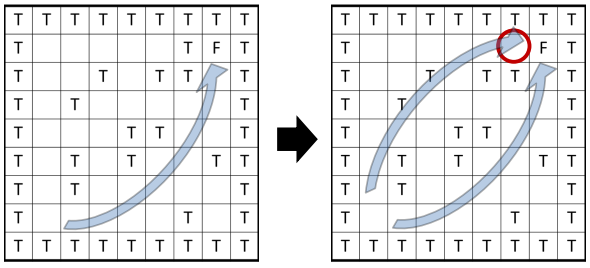
5.2 評価方法とパラメータ設定

評価方法は，餌の獲得までに要したステップ数の平均 (performance) を評価基準とする．ステップ数の平均値は，最短経路の平均ステップ数 (optimum) に近いほど正確に学習していることを示す．また，評価基準は学習回数 50 回ごとの移動平均で示す．

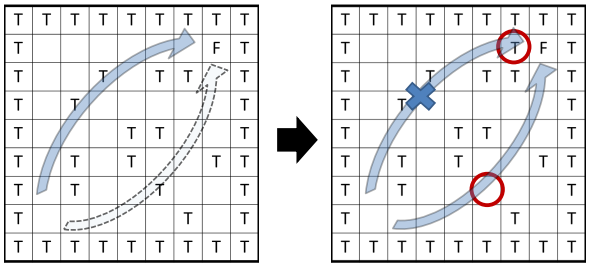
また，XCS ならびに XCSAM のパラメータは先行研究 [2][7] を参考にして次のように設定する．具体的には，$N = 3000$, $P_{\#} = 0.3$, $\beta = 0.2$, $\gamma = 0.7$, $\chi = 0.8$, $\mu = 0.01$, $\theta_{mna} = 8$, $P_{explr} = 1.0$, $\theta_{GA} = 100$, $\epsilon_0 = 1$, $\theta_{del} = 20$ である．ただし，環境変化後においても安定して最適行



a) 環境変化 I



b) 環境変化 II



c) 環境変化 III

図 5 類似環境への環境変化の種類

動を識別するために，$\zeta = 0.9$ と設定する．提案手法のパラメータは，$\theta_{exp} = 30$，$\theta_{eam} = 1.5$ と設定し，その他は XCSAM と同様に設定する．

5.3 実験結果

5.3.1 知識の獲得と再適用（環境変化 I，II）

図 6, 7 は，I)Maze6-a から Maze6 に環境が変化した場合と II)Maze6 から Maze6-a に環境が変化した場合の XCS, XCSAM, XCSL のステップ数の経過を示している．各図は縦軸が平均ステップ数，横軸は学習回数をそれぞれ示している．

図 6 より，環境変化後 (Maze6) では，XCS ならびに XCSAM はステップ数が収束するまでに，それぞれ 5000 回ならびに 3600 回程度の学習回数が必要であることがわかる．しかし，XCSL は 900 回程度の学習回数でステップ数が収束している．よって，XCSL は環境変化前に獲得した行動制御則を変化後の環境に再適用することで，少ない学習回数で環境に適応可能であることがわかる．

図 7 より，環境変化直後 (Maze6-a) の学習回数 10000 回から 10200 回の間では，XCSL のステップ数が最も少ない．これは，XCS や XCSAM では，環境変化前に学習した行動制御則が変化後の環境に正しく適用できず，適切な最適行動を選択していないことが原因である．しかし，XCSL では，環境変化前の行動制御則を用いることで，環境変化後も同様の行動制御則である状態については，適切な最適行動を選択できるためである．

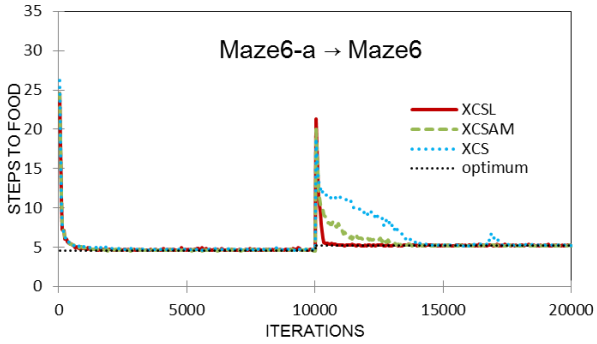


図 6 Maze6-a から Maze6 に環境変化したときの平均ステップ数

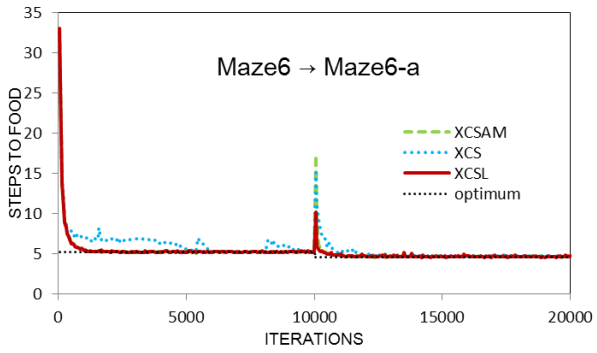


図 7 Maze6 から Maze6-a に環境変化したときの平均ステップ数

5.3.2 知識の獲得状況の影響（環境変化 III）

図 8 は，III)Maze5 から Maze6 に環境が変化した場合の XCS, XCSAM, XCSL のステップ数の経過を示している．図 8 より，環境変化後（Maze6）においては，XCS, XCSAM はステップ数が最適経路のステップ数に到達するまで 4000，2600 回程度の学習回数を必要としているのに対して，XCSL は 1300 回程度の学習回数で収束している．また，環境変化直後に必要とするステップ数が XCS は 19 ステップ，XCSAM は 24 ステップに対して，XCSL は 18 ステップと最も少なかった．

5.4 考察

5.4.1 行動制御則の特徴抽出

XCSL による行動制御則の特徴抽出について議論する．図 9 に環境変化前の Maze6-a および Maze6 において形成されたリンケージ型分類子を示す．図中の各矢印は，リンケージ型分類子によって示される同一の最適行動が適用可能な一連の状態とその行動について示している．また，図 2 に実際に獲得したリンケージ型分類子の構成について獲得例を示す．

ⓒ 2014 Information Processing Society of Japan

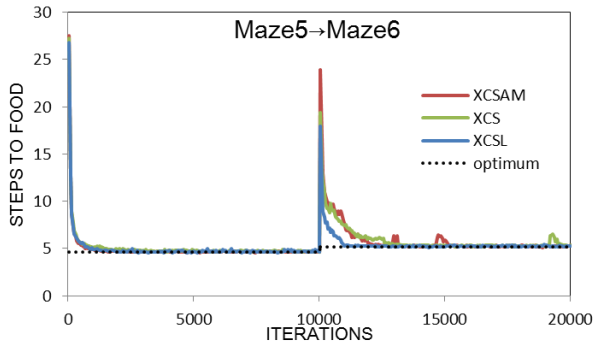


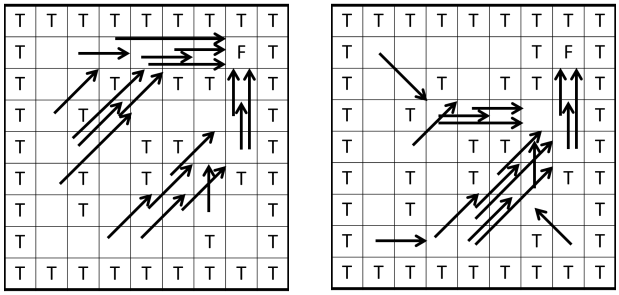
図 8 Maze5 から Maze6 に環境変化したときの平均ステップ数

図 9 より，XCSL は，Maze6-a に関しては，右回りと左回りの 2 通りの経路について，Maze6 では，右回りの経路を示すリンケージ型分類子が適切に形成されていることがわかる．また，Maze6-a ならびに Maze6 において，同一の最適行動をとる迷路の右下の状態群では，類似するリンケージ型分類子が存在する．このことから類似環境の共通する行動制御則の抽出ができていることがわかる．

また，表 2 の獲得例より，条件部 (condition) では 4 つの分類子とも，1, 2 ビット目が “01”，4 ビット目が “1”，14, 15 ビット目が “00” と複数状態において，共通するビットが存在することがわかる．以上より，リンケージ型分類子は，同一の最適行動制御則をもつ分類子を集約することで，逐次的意思決定における行動制御則の特徴を抽出していることがわかる．

5.4.2 環境変化前に獲得した知識の再適用

環境変化前に獲得した知識の再適用能力について議論する．XCS や XCSAM では，環境変化によって分類子の予測値や適合度の値が変化し，正確な分類子として認識されなくなることで，分類子の再適用が困難となる．例えば，Maze6-a から Maze6 に環境が変化した場合，左回りから餌にたどり着く経路は環境変化後の Maze6 では適用すべきでない．その結果，左回りの経路を構成する行動制御則は誤りであり，正しい経路の獲得に向けて再学習しなければならない．この再学習により，正しい経路である右回りの行動制御則も，左回りからの行動価値の伝搬や，左回り右回り共通の分類子によって，正確性を落としてしまい，環境全体の再学習となってしまう．よって，XCS や XCSAM では，各分類子の実行順を考慮していないため，各分類子がそれぞれ正しく再学習されるのを待たねばならない．一方で XCSL では，リンケージ型分類子が複数存在する状態では，リンケージ型分類子によってそのリンケージ型分類子の行動を選択するように圧力が加わり，正確性を落とす分類子によって予測報酬値が減少しても，圧力が減少値を補い適切な行動を選択することを可能とすると考えられる．行動選択時の圧力とは，選択確率を上げるために予測値の大きい分類子で予測報酬値を算出することである．リンケージ型分類子は，リンケージの先端である報酬に最も近い分類子の予測値を用いて予測報酬値を算出する，つまり，リンケージ型分類子を構成している分類子で最大の予測値を持つ分類子で予測報酬値を算出し行動選択を行うため，リンケージ内の各状態においてその行動を選択しやすくなるのである．



a) Maze6-a　　b) Maze6

図 9 リンケージ型分類子（環境変化前）

表 2 リンケージ型分類子の獲得例

condition	*a*	*p*	*exp*	*eam*
01010#0001000001	2	334.4	448	1.00
01#10###00010001	2	485.9	902	1
0101###101###001	2	691.9	2536	1
01#11#0###01000#	2	1000	1344	1

5.4.3 知識獲得状況の影響

XCSL のリンケージ型分類子の獲得状況による，学習能力への影響について議論する．図 10 に 1)Maze6-a で学習した後に Maze6 に環境変化した場合と，2)Maze5 で学習した後に Maze6 に環境変化した場合の環境変化後の環境適応能力の比較と，Maze5 で形成されたリンケージ型分類子を示す．

図 10 より，1)Maze6-a から Maze6 に環境変化した場合の方が 2)Maze5 から Maze6 に環境変化した場合よりも早く最適経路に収束していることが分かる．このことは，リンケージ型分類子の獲得状況が影響していると考えられる．具体的には，Maze6-a と Maze5 の違いは右下に障害物が存在するかしないかであるが，この障害物によって，図 9, 10 から Maze6-a と Maze5 のリンケージ型分類子の獲得状況を比較してわかるように，Maze5 は右回りから餌に到達する経路を取りにくい環境となっているため，右回りの経路を取るリンケージ型分類子の獲得が少なかった．いずれの環境変化においても左回りから餌に到達する経路が環境変化によって絶たれることになるため，右回りの経路のリンケージ型分類子をより多く獲得していることが，環境変化後の適応能力を上げるのである．

次に，XCSAM との性能差の観点から考察する．Maze5 から Maze6 に環境変化した際の XCSAM と XCSL を比較

すると，XCSL の方が早くステップ数が収束している．このことから右回りの経路を取りにくい Maze5 のような環境でも，獲得した少ない知識を有効に活用することで，早く新しい環境に適応できることが明らかとなった．

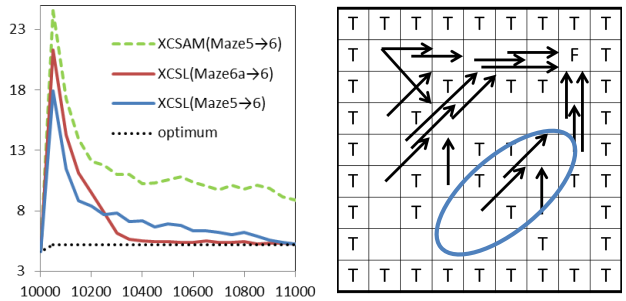


図 10 平均ステップ数の比較と Maze5 のリンケージ型分類子

5.4.4 淘汰を考慮したリンケージの解除

図 11 に，XCSL のリンケージ解除条件に加えて，[P] で淘汰された分類子についてもリンケージを解除する XCSL-R の，環境変化 I におけるステップ数の経過を示す．図 11 より，環境変化直後に必要とするステップ数が XCSL は 21 ステップ，XCSAM は 20 ステップに対して，XCSL-R は 14 ステップと最も少なかった．また，ステップ数の収束に関しては，XCSL と同等の 900 回程度であった．

このことから，淘汰された分類子のリンケージ解除を行うことによって，環境変化直後に必要とするステップ数が 66%に削減されることが明らかとなった．以上から，淘汰された分類子から構成されるリンケージ型分類子の中には，類似環境への適応を妨げるものが存在し，その解除を行うことによって環境変化直後のステップ数を抑えることが可能であることが明らかとなった．

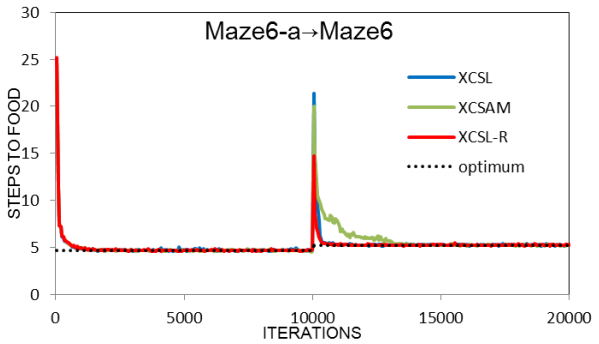


図 11 Maze6a から Maze6 に環境変化したときの平均ステップ数

6. おわりに

本論文では，環境が変化した場合でも再適用可能な汎用的行動規則の獲得手法を提案した．具体的には，最適行動制御則の前後関係を考慮したリンケージ型分類子を導入し，その作成法ならびに解除法を加えたリンケージ型学習分類子システム (XCSL) を提案した．逐次意思決定問題のベンチマーク問題であるロボット制御問題 (Maze 問題) に提案手法を適用したところ，次の知見を得た．1)XCSL は環境が変化した場合でも，従来手法である XCS や XCSAM よりも少ない学習回数で最短経路を学習できることを示した．2)XCSL の類似環境への適応能力は，環境変化後でも変化しない経路に関してより多くのリンケージ型分類子が得られることによって適応能力が高くなることを示した．3) 淘汰も考慮したリンケージの解除を行ったところ環境変化直後に必要とするステップ数が削減されることを示した．

今後は，実問題の適用を目指して，次の実問題を想定した環境において提案手法を拡張する．まず，報酬値や行動部が実数値を扱う実数値環境問題に適用可能な手法を検討する．次に，より複雑な逐次意思決定問題として，不完全知覚問題や雑音を付加した環境において，提案手法の有効性を検証する．

参考文献

[1] M. V. Butz, S. W. Wilson, “An algorithmic description of xcs,” *Journal of Soft Computing*, 6(34):144-153, 2002.
[2] M. V. Butz, D.E. Goldberg and P. L. Lanzi. “Gradient Descent Methods in Learning Classifier Systems: Improving XCS Performance in Multistep Problems,” *Evolutionary Computation*, 9(5):452-473, 2005.
[3] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison Wesley, 1989.
[4] J. H. Holland, “Escaping Brittleness: The Possibilities of General Purpose Learning Algorithms Applied to Parallel Rule-based system,” *Machine Learning*, 2:593-623, 1986.
[5] P. L. Lanzi, “An Analysis of Generalization in the XCS Classifier System,” *Evolutionary Computation Journal*, 7(2):125-149, 1999.
[6] 森本淳，杉本徳和，“強化学習の最近の発展《第 9 回》高次元・実環境における強化学習”，計測と制御, 52(8):742-748, 2013.
[7] M. Nakata, P. L. Lanzi and K. Takadama, “XCS with Adaptive Action Mapping,” *Simulated Evolution and Learning*, The Ninth International Conference on Simulated Evolution And Learning (SEAL 2012), LNCS 7673:138-147, 2012.
[8] N. Sugimoto and J. Morimoto, “Phase-dependent Trajectory Optimization for CPG-based Biped Walking Using Path Integral Reinforcement Learning,” *IEEE-RAS International Conference on Humanoid Robots (Humaniods 2011)*, :255-206, 2011.
[9] N. Sugimoto and J. Morimoto, “Off-line path integral reinforcement learning using stochastic robot dynamics approximated by sparse pseudo-input Gaussian processes: Application to humanoid robot motor learning in the real environment,” *ICRA 2013*, :1311-1316, 2013.
[10] R. S. Sutton, “Learning to Predict by the Methods of Temporal Differences,” *Machine Learning*, 3(1):9-44, 1988.
[11] R. S. Sutton, A. G. Barto, *Reinforcement Learning - An Introduction*, MIT Press, 1998.
[12] S. W. Wilson, “Classifier fitness based on accuracy,” *Evolutionary Computation*, 3(2):149-175, 1995.

ⓒ 2014 Information Processing Society of Japan