Vol.2013-MPS-95 No.12
2013/9/26

複素強化学習を用いた学習分類子システムによる PoMDPs 環境への展開

山﨑大地[†1]　　中田雅也[†1,2]　　高玉圭樹[†1]

本論文では，難解な部分マルコフ決定過程（PoMDPs）環境において最適な方策を獲得するために，複素強化学習（Complex-Valued Reinforcement Learning : CVRL）を用いた学習分類子システム（CVRL-based Classifier System: CVRL-CS）を提案する．計算機実験では，従来手法（Q-Learning と ZCSM）が適用困難な PoMDPs 環境として，a) 状態空間が大きい環境，b) 不完全知覚の特性が異なる環境に提案手法を適用したところ，1) 提案手法は，従来手法よりも少ない学習回数で高い学習性能を実現し，2) 従来手法が学習不可能であるのに対し，初期状態が不完全知覚となる問題においても，提案手法は最適な方策を獲得可能であることを示した．

1. はじめに

環境における適切な行動制御則（方策）の獲得を目的とした強化学習（Reinforcement learning : RL）[1]や学習分類子システム（Learning Classifier System : LCS）[2]は，環境から与えられる報酬を用いて試行錯誤的に適切な方策を学習する．これらの手法を実環境へ適用する場合，環境からの外乱やエージェントの状態知覚誤差により，異なる環境状態を同一状態と誤って知覚する不完全知覚問題[3]が生じる．そのため，不完全知覚問題を有する部分観測マルコフ決定過程（Partially Observable Markov Decision Processes : PoMDPs）を扱う環境への適応が重要な課題である[4].

PoMDPs 環境に適用可能な RL 手法として，複素強化学習（Complex-valued Reinforcement Learning : CVRL）[5]が提案されている．CVRL では複素数化した行動価値によって行動履歴から行動文脈を構築し，不完全知覚状態を特定する．しかし，CVRL は環境との試行錯誤的なやり取りのみで学習を進めるため，報酬獲得に多数の試行が必要となる環境における学習効率が低下するという問題が存在する．一方，PoMDPs 環境に適用可能な LCS としては，進化型メモリベース法を組み込んだ ZCSM（Zeroth level Classifier System with Memory）[6]が主流である．進化型メモリベース法の特徴は，1) 過去の状態行動履歴をメモリとしてルールに付加することで不完全知覚状態を知覚すること，2) 履歴情報（メモリ）とルールの正しい組み合わせを進化的に探索することである．しかし，A) 複数の不完全知覚状態を有する環境ではメモリサイズが膨大となること，それに伴って B) メモリと分類子の組み合わせ数が膨大となることで探索効率が低下することといった問題を有する．

そこで本稿では，CVRL を学習分類子システムに組み込んだ CVRL-based Classifier System（CVRL-CS）を提案する．提案手法は，CVRL の問題点を，1) 方策（分類子）の探索効率を進化計算によって向上させることで克服する．

†1 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
†2 日本学術振興会　特別研究員（DC1）
Research Fellow of Japan Society for the Promotion of Science（DC1）

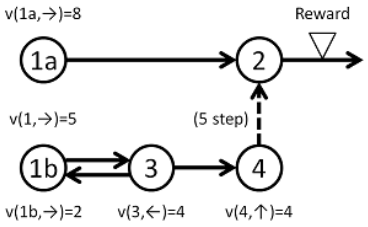

図 1 Type 1 の混同

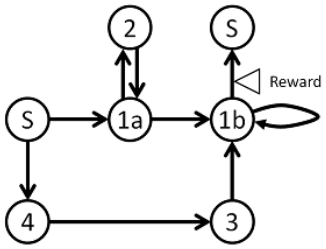

図 2 Type 2 の混同

さらに，ZCSM の問題点に対し，2) 複素数による文脈依存型学習を行うことで，メモリを必要とせず不完全状態を知覚する．提案手法の有効性を検証するため，仮想迷路環境（Woods 問題）での計算機実験から従来手法と学習性能を比較する．

2. 部分マルコフ決定過程　(PoMDPs)

PoMDPs 環境下では，異なる方策を獲得すべき別々の状態を同一の状態として知覚する問題（不完全知覚問題）によって，適切な方策の獲得が困難となる．宮崎らは不完全知覚によって，エージェントが自身の状態を誤認してしまう状況を混同と定義し，2 種類の混同を定義している[4].

2.1 Type 1 の混同

Type 1 の混同では，行動価値が異なる状態を同一の状態として知覚する．例えば，図 1 に示す環境において，エージェントは状態 1a と 1b を同一の状態として知覚する．状態 s から上下左右に移動する行動価値を$v(s,\{\uparrow,\downarrow,\leftarrow,\rightarrow\})$と表現し，各状態の行動価値を「10－報酬までの最短ステップ数」と推定する場合，$v(1a,\rightarrow)=8$，$v(1b,\rightarrow)=2$となる．そのため，エージェントは不完全知覚によって$v(1,\rightarrow)=5$と推定する．同様に $v(3,\rightarrow)=3$となるが，状

態 1 の価値を考慮すると$v(3,\leftarrow)=4$と推定される．このように，報酬から遠ざかる行動価値を誤って高く見積もることで，適切な方策を獲得できなくなる問題が生じる．

2.2 Type 2 の混同

Type 2 の混同では，適切な行動が異なる状態を同一の状態として知覚する．例えば，図 2 に示す環境においてエージェントは状態 1a と 1b を同一の状態として知覚する．ここで初期状態を状態 S とすると，最短で報酬を獲得するために状態 1a で右に移動する行動を学習した場合，状態 1b では報酬を獲得する行動を選択できなくなる．同様に，状態 1b で上に移動する行動を学習した場合，状態 1a と状態 2 を往復し続けることとなる．このように，適切な行動が異なる 2 つの状態を同一の状態として知覚することで，適切な方策を獲得できなくなる問題が生じる．

3. 複素強化学習

CVRL の中でも代表的な手法である $\dot{Q}$-learning は，PoMDPs 環境下の RL 手法として，価値関数を複素数化することで方策の文脈を表現可能にした手法である[5][7]．$\dot{Q}$–learning では，エージェントが環境から入力された状態sと複素行動価値$\dot{Q}(s,a)$から行動aを選択し，得られた報酬を基に行動価値を更新することで適切な方策を学習する．$\dot{Q}(s,a)$ではその価値の大きさを絶対値で表し，時系列上での文脈情報を位相で表す．選択する行動価値の位相を時間とともに回転させることで，時系列に対する方策の文脈を表現することができる．

3.1.1 複素行動価値の更新

$\dot{Q}$-learning において環境から与えられた報酬r_{t+1}に対する行動価値関数の更新式は，以下の式(1)，(2)で表される．ここで $\dot{I}$ は時刻tにおける内部参照値， $\bar{\dot{I}}$ は $\dot{I}$ の複素共役，$Re[\cdot]$は複素数の実部を表す．また，$\dot{\beta}$は時間経過に対する位相回転量のパラメータであり，パラメータαおよびγは学習率および割引率を表す．$\dot{Q}$ -learning における行動価値関数の更新では，次状態での行動価値$\dot{Q}^{(t)}{}_{max}$から位相が $\dot{\beta}^{k+1}$ だけ回転した値に近づくように行動価値関数が更新される．$k=0,1,\ldots,Ne-1$であり，Neは何ステップ前の価値関数を参照するか決定するトレース数と呼ばれるパラメータである．

$$\dot{Q}(s_{t-k},a_{t-k}) \leftarrow (1-\alpha)\dot{Q}(s_{t-k},a_{t-k}) + \alpha\left(r_{t+1}+\gamma\dot{Q}^{(t)}{}_{max}\right)\dot{\beta}^{k+1} \quad (1)$$

$$\dot{Q}^{(t)}{}_{max} = \dot{Q}\left(s_{t+1}, \underset{a}{\operatorname{argmax}}\left(Re\left[\dot{Q}(s_t,a)\bar{\dot{I}}_t\right]\right)\right) \quad (2)$$

3.1.2 複素行動価値による行動選択

$\dot{Q}$ -learning における行動選択では，1) 複素行動価値の絶対値の大きさと 2) 複素行動価値の位相と内部参照値の位相の近さの観点から行動を決定する．絶対値の大きさは将来的な期待収益が大きい行動を選択する指針となる一方，内部参照値との位相の近さは時系列の文脈に沿った行動を選択する指針となる．

内部参照値は式(3), (4)で表される．内部参照値の位相は，直前に選択した複素行動価値に対して，式(1)によって回転した位相$\dot{\beta}$と逆の回転を加えたものとなる．式(3)によって更新された複素行動価値が収束した場合，次状態の行動価値と内部参照値の位相が一致することから，このような内部参照値によって時系列の文脈が表現可能となる．なお，初期状態においては直前の複素行動価値が存在しないため，式(4)に示す内部参照値を用いる．

$$\dot{I}_{-1} = \dot{Q}\left(s_0, \underset{a}{\operatorname{argmax}}\left(\dot{Q}(s_0,a)\right)\right) \quad (3)$$

$$\dot{I}_t = \dot{Q}(s_t,a_t)/\dot{\beta} \quad (4)$$

行動選択に関する 2 つの観点の両方を考慮した行動選択方策として，式(5)に示すような複素行動価値のためのボルツマン選択法が提案されている．ここで$\pi_{i_{t-1}}(s_t,a')$は時刻tの状態s_tにおいて行動a'を選択する確率である．また，$T(>0)$は温度定数と呼ばれるパラメータである．式(5)中の$Re\left[\dot{Q}(s_t,a')\overline{\dot{I}_{t-1}}\right]$という表現は，式(2)において $\dot{Q}^{(t)}{}_{max}$を決定する際にも用いられているように，内部参照値に近く，かつ絶対値の大きい行動価値ほど大きな値となる．

$$\pi_{i_{t-1}}(s_t,a') = \frac{\exp\left(Re\left[\dot{Q}(s_t,a')\overline{\dot{I}_{t-1}}\right]/T\right)}{\sum_a \exp\left(Re\left[\dot{Q}(s_t,a)\overline{\dot{I}_{t-1}}\right]/T\right)} \quad (5)$$

4. ZCSM

4.1 ZCSM の概要

ZCSM[6]は，学習と進化の 2 つの概念を取り入れた環境適応システムである ZCS[8]を PoMDPs 環境下に適用するために，メモリベース法として内部レジスタを組み込んだシステムである．ZCSM は条件部と行動部からなる IF-THEN ルール（分類子）とルール集合（Population : [P]），0,1 のビット列から形成される b ビットの内部レジスタを持つ．を持つ．ルールを環境に実行したことで得られる報酬を用いて，RL によって強度値を更新することで，報酬を最大化するルールを学習する．さらに，遺伝的アルゴリズム（Genetic Algorithm : GA）[9]を用いて分類子を進化させることで, 膨大な状態行動空間から最適なルールを探索する．

4.2 分類子

ZCSM の分類子は条件部と行動部，内部条件部，内部行動部，それらの価値となる強度値（Strength : S）から構成される．各部位は 0,1 のビット列によって表現される．行動部を除く部位については任意の値を意味する#（don't care）記号を組み込むことで，汎用的な分類子を表現できる．

4.3 メカニズム

ZCSM のメカニズムは，（1）実行部，（2）強化部，（3）発見部から構成される．

4.3.1 実行部

実行部では，環境から入力された状態において，適切な行動を出力するまでの処理を行う．環境から入力された状態は 0,1 のビット列から形成される．ここで，環境入力と条件部が照合し，かつ，内部レジスタと内部条件部が照合した分類子を[P]から抽出し，照合集合（Match Set : [M]）を形成する. ここで, 1) [M]が空である場合, もしくは 2) [M]内の分類子の強度値の合計が，[P]内の分類子の強度値の合計にパラメータφを掛けた値よりも低い場合，入力状態に一致する条件部を持つ分類子を生成する．この処理を被覆（Covering）と呼ぶ．被覆によって生成される分類子の条件部・内部条件部の各ビットは確率 $P_{\#}$で#に置き換えられ，行動部・内部行動部はランダムに設定される．また，強度値は初期値 S_0 に設定される．

次に，[M]内の分類子の強度値を選択確率としたルーレット選択により分類子を 1 つ選択し，[M]から選択分類子と同じ行動部・内部行動部を持つ分類子を抽出して行動集合（Action Set : [A]）を形成する．その後，行動を環境に対して実行し、内部レジスタを内部行動の値に置き換える．ただし，内部行動が #であるビットは無視される．この一連の処理の流れを 1 ステップと呼ぶ.

4.3.2 強化部

強化部では，実行部を実行後，環境から得られた報酬rに基づき, 式(6)から 1 ステップ前の行動集合$[A_{-1}]$内の各分類子cl_iの強度S_jを更新する． $[A']$は次ステップにおいて，分類子の強度値の合計が最大となる行動による行動集合，$|A_{-1}|$は 1 ステップ前の行動集合内の分類子数を意味する．パラメータα，γは学習率，割引率である．また，1 ステップ前の照合集合内に含まれているが行動集合内に含まれていない各分類子については, その強度にパラメータτが掛けられることで強度が減衰する．

$$S_j \leftarrow S_j + \alpha\left(\left(r + \gamma \sum_{cl_i \in [A']} S_i\right)\Big/_{|A_{-1}|} - S_j\right) \tag{6}$$

4.3.3 発見部

発見部では，GA を用いて[P]内の分類子を進化させることで，適切なルールを探索する．GA は，エージェントの学習が終了するたびパラメータρの確率で実行される．GA が実行される場合，[P]内の分類子の強度値からルーレット選択によって親個体となる分類子が 2 つ選択される. 次に，子個体として，各親個体の分類子と同様のビット列および強度値を持つ分類子を 2 つ生成し，交叉（crossover）および突然変異（mutation）を適用する．交叉はパラメータ χ の確率で実行され，2 つの個体の条件部，内部条件部および内部行動部の一部を交換することで新たな個体を生成する．突然変異は，各個体の持つ各ビットについて，確率μでランダムなビットに変化させる．最後に，各行動の選択率を保持するために親個体および子個体の価値を半分にし，生成した子個体を[P]に追加する．この際，[P]の分類子数が分類子上限数 N を超えた場合，分類子の強度値を選択確率として，削除する分類子を決定し削除する．

5. 提案手法

Q̇-Learning はメモリを使用せず行動価値と内部参照値のみから方策の文脈を表現することが可能であるが，報酬獲得機会が少ない環境においては学習効率が低下するという問題が存在する．一方．ZCSM は，適切な方策を進化的に探索することで Q̇–Learning の問題を克服できるメカニズムを有するが，必要とするメモリのビット長によっては探索空間が増大するため, 不完全知覚状態が特定困難となる．

そこで本稿では，Q̇ -Learning を学習分類子システムに組み込んだ Complex-valued Reinforcement Learning–based Classifier System（CVRL-CS）を提案する．提案手法は，Q̇-Learning の問題点を，1) 方策（分類子）の探索効率を進化計算によって向上させることで克服する. さらに, ZCSM の問題点に対し, 2) 複素数による文脈依存型学習を行うことで，メモリを必要とせず不完全状態を知覚可能である．

5.1 CVRL-CS における分類子

CVRL-CS で扱う分類子は，条件部と行動部，Q̇ -Learning における複素行動価値から構成される．しかし, #を条件部に含む汎用的な分類子は，誤った文脈を形成するため，CVRL-CS における分類子は#を用いない. また, Q̇ –Learning の内部参照値とは異なり, CVRL-CS における内部参照値は式(7)，(8)により示される．

$$\dot{I}_t = \sum\nolimits_{cl_i \in [M_t]} \dot{S}_i / \dot{\beta} \tag{7}$$

$$\dot{I}_{-1} = \underset{cl_i \in [M_0]}{\operatorname{argmax}}\left(\left|\dot{S}_i\right|\right) \tag{8}$$

5.2 CVRL-CS のメカニズム

5.2.1 実行部

CVRL-CS では ZCSM と同様に [M]を形成した後，環境に出力する行動を式(9)の確率に従って選択する．ただし，式中の$[M]|a$は，[M]内で行動部がaである分類子の集合を表し，空の場合には$\sum_{cl_i \in [M]|a} Re\left[\dot{S}_i \overline{\dot{I}_{t-1}}\right]$を 0 として扱う．

$$\pi_{i_{t-1}}(a') = \frac{\exp\left(\sum_{cl_i \in [M]|a'} Re\left[\dot{S}_i \overline{\dot{I}_{t-1}}\right]/T\right)}{\sum_a \exp\left(\sum_{cl_j \in [M]|a} Re\left[\dot{S}_j \overline{\dot{I}_{t-1}}\right]/T\right)} \tag{9}$$

Covering の処理は，入力と条件部が一致する分類子が分類子集合内に存在しない場合，もしくは選択された行動と一致する行動部を持つ分類子が[M]に存在しない場合に実行する．後者の条件によって被覆が実行された場合，生成される分類子の行動部は選択された行動と同一となる．

5.2.2 強化部

環境から報酬rについて，kステップ前までの行動集合$[A_{-k}]$内にある各分類子の価値$\dot{S}_j$を式(10), (11)のように更新

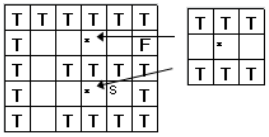

図 3 Woods 問題の例

する．ここでは適格度トレースを ZCS の更新式に適用しているため，Q̇ -Learning と異なり$k = 1,2, \ldots, Ne$としている．

$$\dot{S}_j \leftarrow (1-\alpha)\dot{S}_j + \alpha\left(\left(r + \gamma\dot{S}^{(t)}{}_{max}\right)\Big/_{|A_{-k}|}\right)\dot{\beta}^k \tag{10}$$

$$\dot{S}^{(t)}{}_{max} = \max_a \sum_{cl_i \in [A]|a} Re\left[\dot{S}_i\overline{\dot{I}_{t-1}}\right] \tag{11}$$

また，CVRL-CS では，内部参照値の値によってはルールの価値が減衰してもルールの選択率が増加する場合があるため，パラメータτによる価値の減衰処理を行わない．

5.2.3 発見部

CVRL-CS は，ZCSM と同様に GA より分類子を進化させるが，行動系列を考慮した上で環境に対して適切なルールを持つ個体を親として選択するために，親の選択はその時刻での内部参照値を用いて[A]から行う．CVRL-CS において GA が実行される条件は，[A]内の各分類子が GA の発動対象となってから経過したステップ数の平均値がパラメータθ_{GA}の値を上回る場合である．式(12)に示される親個体選択確率を基に親個体を選択後，親個体と同一の条件部と行動部を持つ子個体を生成する． CVRL-CS では，誤った行動文脈を形成する個体を生成するため交叉および突然変異を適用しない．子個体を[P]に追加する際，[P]の個体数が最大数 N を超えた場合には，[A]内から式(12)の逆数に従う確率で削除個体を選択する．ただし，[A]内の個体数が 2 つ以下の場合は選択圧を保持するため，[P]内から価値の絶対値の逆数を選択確率として分類子を削除する．

$$P_{i_{t-1}}(cl_i) = \frac{\exp\left(Re\left[\dot{S}_i\overline{\dot{I}_{t-1}}\right]/T\right)}{\sum_a \exp\left(\sum_{cl_j \in [A]} Re\left[\dot{S}_j\overline{\dot{I}_{t-1}}\right]/T\right)} \tag{12}$$

6. 評価問題

6.1 Woods 問題

CVRL-CS の有効性を評価するため，PoMDPs 環境における一般的なベンチマーク問題である Woods 問題を評価問題として用いる． Woods 問題は迷路問題の一種であり，学習エージェントは，図 3 に示すような格子状のフィールド中で初期位置 S から食料（Food : "F"）に到達することを目的としている．フィールドは障害物（Obstacle : "T"），通路（Empty position : " "），F で構成される．エージェントは現在位置から 8 近傍を状態として知覚可能であり，その 8 近傍に移動可能である．ただし，T へ移動した場合は現在位置に留まる．エージェントが F に到達した場合にのみ報酬

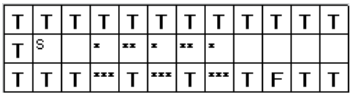

図 4 Type1-Small

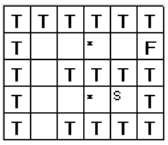

図 5 Type2-Small

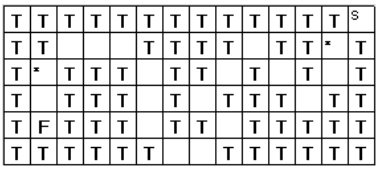

図 6 Type1-Large

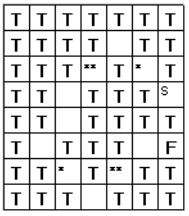

図 7 Type1-2

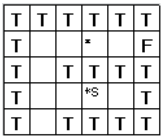

図 8 Type2'

値rが与えられる．

6.2 問題環境

図 4，図 5 に，一般的な PoMDPs 環境として Type 1 および Type 2 に分類される Woods 問題（Type1-Small, Type2-Small）の各フィールドを示す．以降，図中の*,**,***は不完全知覚となる地点を示している．続いて，より広大な状態行動空間を扱う問題（Type1-Large）のフィールドを図 6 に示す． 最後に，Type 1 と Type 2 の混同が複数混在する PoMDPs 環境である問題（Type1-2，Type2'）のフィールドを 図 7, 図 8 に示す. Type1-2 では, *の地点では Type1 の不完全知覚が発生し，**の地点では Type2 の不完全知覚が発生する．また，Type2'は図 5 のフィールドと同じ構造をもつが，初期位置が不完全知覚状態となる点で異なる．初期状態では時系列による文脈を利用できないため，最短経路で F に到達する方策を獲得することが困難となる．

7. 実験

提案手法の有効性を評価するために，第 6 章で説明した 1) 通常の PoMDPs 環境（図 4，図 5），2) 広大な PoMDPs 環境（図 6）ならびに 3) Type 1 と Type 2 の混同が複数混

表 1　各手法におけるパラメータ

Parameter	CVRL-CS	ZCSM	$\dot{Q}$-learning
α	0.1	0.1	0.1
γ	0.9	0.9	0.9
r	100	100	100
S_0	0	20	0
θ_{GA}	20	-	-
b	-	1	-
φ	-	0.5	-
ρ	-	0.25	-
μ	-	0.5	-
χ	-	0.002	-
$P_{\#}$	-	0.33	-
N	800	800	-
τ	-	0.9	-
Ne	2	-	2
T	0.1	-	1

表 2　各問題における基本位相

Parameter	β
Type1-Small	$\exp(i\pi/1)$
Type2-Small	$\exp(i\pi/4)$
Type1-Large	$\exp(i\pi/4)$
Type1-2	$\exp(i\pi/4)$
Type2'	$\exp(i\pi/4)$

在する PoMDPs 環境（図 7，図 8）における各 Woods 問題に提案手法を適用する. また, 後述する評価基準を用いて，従来手法である $\dot{Q}$-Learning と ZCSM と学習性能を比較する.

7.1　評価基準とパラメータ設定

評価基準としては，エージェントが F に到達するか 500 ステップ経過までのステップ数を比較する．低いステップ数であるほど，フィールドにおいて適切な方策を獲得できていることを意味する. 1 試行あたり 10000 回学習を行い，評価基準は学習回数 200 回ごとの移動平均で示す．また，評価基準は，50 試行での平均をとる．

CVRL-CS，ZCSM および $\dot{Q}$-learning で用いる各パラメータ設定を表 1 に示す．各手法における α,γ,r および $\dot{Q}$-learning の各パラメータについては文献[5]，ZCSM の各パラメータについては文献[6]と同様の設定としている．加えて, CVRL-CS および $\dot{Q}$-learning では問題ごとに異なる基本位相$\dot{\beta}$を設定する（表 2）．これらの手法は 1 ステップごとに基本位相だけ回転させた行動価値を学習するため，Type1 において同一の行動を取るべき地点では行動価値の位相が同一になるように，Type2 において異なる行動を取るべき地点では行動価値の位相が真逆となるように，基本位相を設定する．Type1-2 に関しては， Type1 と Type2 の不完全知覚が混在するため，各不完全知覚の行動価値の差が 90° となるように基本位相を設定している．

7.2　標準的な PoMDPs 環境

図 9，図 10 に Type1-Small と Type2-Small における各手法のステップ数を示す. CVRL-CS は両フィールドにおいて $\dot{Q}$ -Learning と同程度の学習回数で，ステップ数が最短経路(optimum)に収束している．CVRL-CS と ZCSM を比較すると，Type1-Small では CVRL-CS の収束結果がわずかに劣るが，Type2-Small においては ZCSM によって適切な方策が

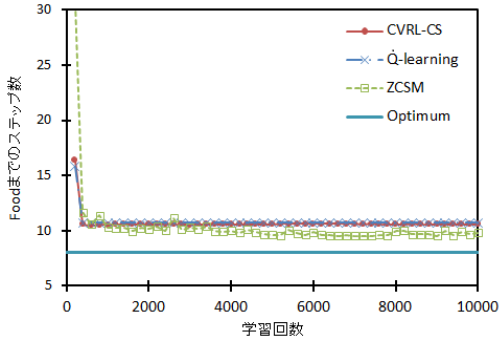

図 9 Type1-Small の学習結果

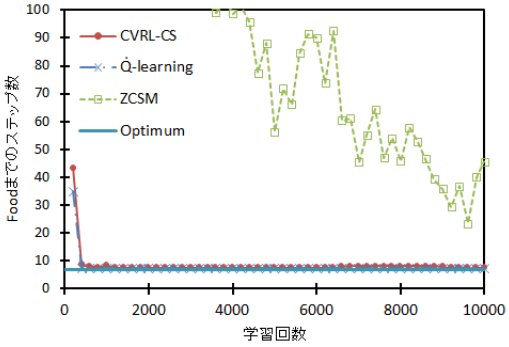

図 10 Type2-Small の学習結果

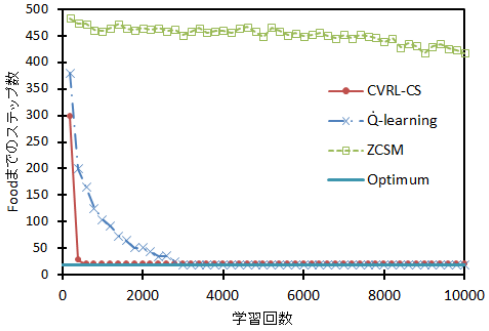

図 11 Type1-Large の学習結果

獲得できない結果となった.

7.3　広大な PoMDPs 環境

図 11 に Type1-Large における各手法のステップ数を示す. ZCSM は大きい状態空間に対応できず，最適な方策を獲得できていない. また, CVRL-CS は, $\dot{Q}$ -learning と比較して，少ない学習回数でステップ数が収束していることがわかる. これは学習初期において，価値の低い不適切なルールを淘汰することで，不適切なルールの選択率を減少させているためである．これによって，相対的に選択すべきルールの選択率が増加することで, ステップ数の収束が促進される.

7.4　Type 1 と Type 2 の混同が複数混在する PoMDPs 環境

図 12，図 13 に，Type1-2，Type2'における各手法のステップ数を示す．図より，ZCSM は両フィールドにおいて学習に失敗している． 図 12 より，異なる Type の不完全知覚が混在しており，適切な基本位相が設定できない環境においても, CVRL-CS は最も少ない学習回数でステップ数が収束している.

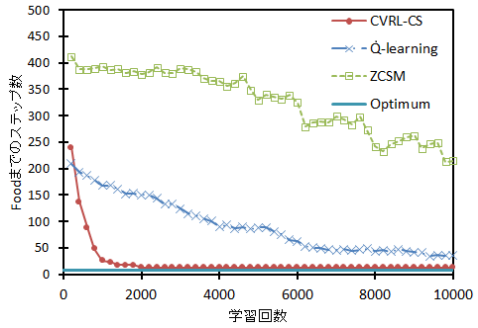

図 12 Type1-2 の学習結果

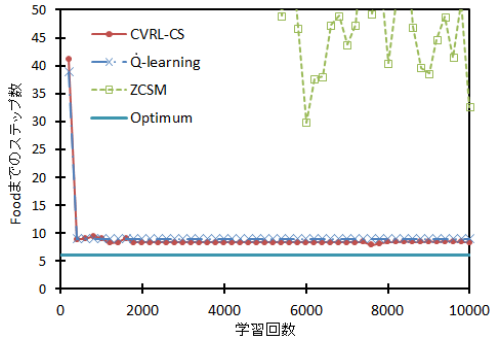

図 13 Type2'の学習結果

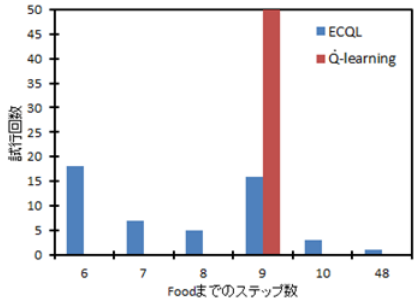

図 14 Type2'における最終的なステップ数の分布

Type2'では，CVRL-CS と Q̇ -learning では同程度の学習回数でステップ数が収束している．ここで各試行の学習終了時のステップ数を図 14 に示す．Q̇-learning は全試行において最適な方策を獲得できていないが，CVRL-CS は 18/50 の試行で最適な方策を獲得していることが分かる．Q̇-learning における初期内部参照値は，初期状態 S において最も絶対値の大きい複素行動価値であり，これは報酬に近い地点の行動と同じ行動を持つ価値となる．Type2'のように初期状態が Type 2 の不完全知覚である環境では，行動系列を表現できないために最適な方策を獲得できない場合がある．一方，CVRL-CS では最も大きな絶対値を持つルールの価値が初期内部参照値となるため，初期状態において適切な行動を持つルールの価値が最大であれば，価値の合計値に関わらず最適な方策を獲得することができる．

8. まとめ

本論文では，複雑な PoMDPs 環境下においても最適な方策を獲得するために，複素強化学習を組み込んだ学習分類子システム（CVRL-CS）を提案した．提案手法は，行動履歴から不完全知覚状態を知覚し，最適な方策を進化的に効率よく探索可能である．1) 通常の PoMDPs 環境，2) 広大な PoMDPs 環境ならびに 3) Type 1 と Type 2 の混同が複数混在する PoMDPs 環境を表現する Woods 問題について提案手法を適用したところ，次の知見を得た．まず，1) 提案手法は，従来手法（Q̇ -learning と ZCSM）よりも少ない学習回数で高い学習性能を実現し，2) 従来手法では学習不可能である，不完全知覚に対して適切な基本位相が設定できない環境においても学習が可能であり，3) 従来手法が最適な方策を獲得不可能な，初期状態が不完全知覚となる問題においても，提案手法は最適な方策を獲得可能であることを明らかにした．

今後の課題としては，1) 行動文脈を考慮した分類子の遺伝的操作法（交叉と突然変異）を考案することで，より効率的な方策の探索法を構築する．また，2) 様々な問題環境に対して適応的に基本位相を設定するアルゴリズムの実装によって，基本位相に対する頑強性の向上を目指す．

参考文献

1) Sutton, R. and Barto, A.（著），三上 貞芳，皆川 雅章（訳）：強化学習，森北出版 (2000).
2) Holland, J.H.: Escaping brittleness: the Possibilities of General-purpose Learning Algorithms Applied to Parallel rule-based systems, in Michalski, R., Carbonell, J. and Mitchell, T., Eds., Machine Learning: An Artificial Intelligence Approach, Vol. 2, pp.593-623, Morgan Kaufmann (1986)
3) Chrisman, L. : Reinforcement Learnng with perceptual aliasing: The Perceptual Distinctions Approach, Proc. of the 10th National Conference on Artificial Intelligence, pp.183-188 (1992).
4) 宮崎 和光，荒井 幸代，小林 重信 : POMDPs 環境下での決定的政策の学習, POMDPs 環境下での決定的政策の学習，人工知能学会誌, Vol. 14, No. 1, pp.148-156 (1999).
5) Hamagami, T., Shibuya, T., and Shimada, S. : Complex-Valued Reinforcement Learning, Proc. IEEE International Conference on the Systems, Man and Cybernetics 2006, Vol. 5, pp.4175-4179 (2006).
6) Cliff, D., Ross, S. : Adding Temporary Memory to ZCS, Adaptive Behavior, Vol. 3, No. 2, pp.101-150 (1995).
7) 澁谷 長史，濵上 知樹 : 複素数で表現された行動価値を用いる Q-learning，電子情報通信学会論文誌, Vol. J91-D, No.5, pp.1286-1295 (2008).
8) Wilson, S. W. : ZCS: A zeroth level classifier system, Evolutionary Computation, Vol. 2, No. 1, pp.1-18 (1994).
9) Goldberg, D. E. : Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley (1989).