マルチエージェント強化学習のための 局所的政策共有システム —学習分類子システムからのアプローチ*

井上 寛康^{†‡}・下原 勝憲^{†‡}・片井 修[‡]

Local Policy-sharing Systems for Multi-agent Reinforcement Learning — An Approach from the Learning Classifier System*

Hiroyasu INOUE^{†‡}, Katsunori SHIMOHARA^{†‡} and Osamu KATAI[‡]

Multi-agent reinforcement learning features many problems, one of these being state-space explosion based on combinations of policies that each agent has. Generally, if agents can share their policies, they can effectively search their enormous state space; that, however, simultaneously produces a risk of local optimality. Hence we propose a novel policy-sharing system based on the Learning Classifier System, on which agents locally share their policies. The aim of this system is to decrease the probability of falls into local optimality, and to effectively reduce state space by policy sharing. To verify the above, we use simplified soccer, which has discrete space and time.

1. 序論

本論文では空間内を自律的に移動するエージェントの集団（以下、マルチエージェント）における強化学習を取り扱う。マルチエージェント強化学習には解決すべき問題が多くあることが知られており[1], それらには例えば, 不完全知覚問題, 同時学習問題や報酬分配問題などがある。これらのほかに, おのおののエージェントが持つ政策の組合せにより発生する状態空間の爆発があり, 本論文ではこの問題を扱う。一般的にエージェント間で政策の共有が可能であれば, この膨大な空間を効果的に探索できるとされている[2]。しかしながら, エージェント全体（以下, グローバル）で政策を共有することで各エージェントが同じ政策となることは, 探索する状態空間を限定し, より最適な政策の獲得を阻害しうる。すなわち局所最適解に陥りうる。グローバルな政策共有によって局所最適解に陥るかどうかは, エージェントに与えられた機能的制約や環境から決まると考えられるが, その因果関係は事前にはわからず, グローバルな政策共

有は常に有効とはいえない。本論文ではこの問題に焦点をあてる。

マルチエージェント強化学習に関する先行研究は数多く存在する[1]が, マルチエージェント強化学習においてグローバルな政策共有を行った研究は多くはない。それらは, 全エージェントで一つの共有された政策を用いて学習する研究, および全エージェントの政策を平均化しながら学習する研究[2]や, エージェント間で政策を交換する研究[3]などである。これらの研究はグローバルな政策共有の有効性を示すにとどまっており, グローバルな政策共有によって局所最適解に陥りうることについては触れていない。また, その欠点を補う方法についての研究はほとんどされていない。

以上を踏まえ本論文では, グローバルではなく, 一部のエージェント（以下, ローカル）で政策共有を行うシステムを学習分類子システムに基づいて提案する。これにより, グローバルな政策共有のように局所最適解に陥る可能性を下げ, かつ政策共有による状態空間の絞り込みの効果を得ることを目指す。この確認のため, エージェントに与えられた機能的制約や環境によって, グローバルな政策共有が実際に局所最適解に陥ることを示し, ローカルな政策共有がそれを回避することを示す。この検証は, 離散空間・離散時間などの単純化を施したサッカーに基づく実験により行う。

本論文は以下のように構成される。2章ではグローバルおよびローカルな政策共有システムについて述べる。

* 原稿受付 2005年3月10日

† ATR ネットワーク情報学研究所 ATR Network Informatics Laboratory; 2-2-2 Keihanna Science City, Kyoto 619-0288, JAPAN

‡ 京都大学大学院 情報学研究科 Graduate School of Informatics, Kyoto University; Yoshida-Honmachi, Sakyo ward, Kyoto city, Kyoto 606-8501, JAPAN

Key Words: multi-agent system, reinforcement learning, policy sharing, learning classifier system, role.

3章では本論文で用いた単純化したサッカー, 4章では実験の設定について述べる. 5章では実験の結果を示し, 6章でその結果を吟味する. 最後に7章で結論を述べる.

2. 政策共有システム

2.1 ベースとなるシステム

個々のエージェントが強化学習により獲得した政策を(グローバルあるいはローカルな)エージェント間で共有し, 各エージェントがそれを自身に反映する政策共有システムを提案する. そのようなシステムは一つに決まるものではないため, 本論文では一般的と思われる一つのシステムを提案する.

本論文では強化学習を行うシステムとして, 学習分類子システム (Learning Classifier System (LCS)) [4]を採用する. これは, 条件部と行動部からなるルール(分類子)が複数集まったルール群(政策に相当する)を持ち, それらルールが環境から与えられる報酬をより多く得るために競い合うことで, 対象となるシステムを制御するシステムである. このLCSを用いる理由は, 存在可能なルールの一部を持ち, 動的にルールの増減を行うことを前提としたシステムであるため, 後述する政策の共有方法と整合すると考えるためである. 本論文では最も基本的なLCSとして提案されているZeroth level Classifier System(ZCS)[5]を採用する. 従来のLCSにはメモリの役割をするメッセージリストがあるが, ZCSにはない.

本節以降では, ZCSの持つ分類子をルール, そしてZCSによる政策を特にルール群と記述する. また, 各エージェントが持つルール群を合わせてルール群集合と記述する. さらに, 政策共有をルール共有と記述する.

以下で簡単にZCSを, ルールの形式, 実行部, 強化部, 発見部に分けて説明する. 本論文ではオリジナルのZCSを修正して用いているが, 具体的にはそのつど説明する.

2.1.1 ルールの形式

ルールは条件部と行動部の対からなる. 条件部は離散値あるいは任意の値を示すワイルドカードによって表現される. たとえば1, #, 0などは三つの条件からなる条件部の例である. 行動部は任意の表現でよい. ルールはほかに強度を持つ. 以下では, ある*i*番目のルールを r_i とし, その強度を s_i とする.

ルール群には最大のルール数(N)の制限があり, 初期状態では条件部と行動部をランダムに生成した N 個のルールを持つ. その際の s_i の初期値を s_0 とする.

2.1.2 実行部

実行部はルール群を参照し, 状態に対して適切な行動を選択し, 出力する過程を担う. 環境から状態が入力されると, ルール群に対して条件部と状態の照合が行われる. その照合では, ワイルドカード以外は全く一致するルールかどうか調べられる. 照合で一致したルールは

照合群を形成する. なお, 照合時にルール群に一致するルールが存在しない場合, 入力された状態に一致するルールを新たに生成する被覆メカニズムが実行される. この被覆メカニズムでは, 入力された状態に一致するように条件部が作成されるが, そのとき各条件にワイルドカードをある確率($P_{\#}$)で含むように作成される.

照合群内のルールの行動部には複数の行動が含まれて競合することがある. これを解消するために, 競合する行動による強度を連ねた強度配列が作成される. 強度配列における各行動 a_i の強度は次のように計算される.

$$S(a_i) = \sum_{r_k \in [M]|a_i} S_k \quad (1)$$

ただし, $[M]|a_i$ は照合群内の行動部に行動 a_i をもつルールの集合を表す. この強度配列に基づいてルーレット選択により行動が選ばれる. すなわち, ある行動が選ばれる確率 $P(a_i)$ は次のように計算される.

$$P(a_i) = \frac{S(a_i)}{\sum_{k \in \text{Action}} S(a_k)} \quad (2)$$

ただし, Actionは行動の集合である. 選択された行動を持つルールにより行動群が形成される. そして, その行動が実行されることで実行部の一連の処理が終了する. この一連の処理をステップとよぶ.

2.1.3 強化部

強化部は各ステップにおける実行部の処理が完了した後に, 行動群に含まれるルールの強度を環境から得られる報酬に基づいて更新する. オリジナルのZCSではルールの強化をバケツリレーアルゴリズムで行っているが, 本論文ではProfit Sharingアルゴリズム [6]を用いる. これは, マルチエージェント強化学習に対してはProfit Sharingアルゴリズムの優位性が示唆されている [7]ためである. そのProfit Sharingアルゴリズムの報酬関数においては等比減少関数の有効性が指摘されている [8]ため, これを採用する. すなわち, 報酬の t ステップ前に行動群に表れたルール r_i の強度 s_i は

$$s_i \leftarrow s_i + \beta(r^t - s_i) \quad (3)$$

となる. ただし, r は上述の等比減少関数の比であり, β は学習率である. なお左矢印は代入を意味する.

2.1.4 発見部

オリジナルのZCSでは, 膨大なルールの中から有効なルールを作り出すために, 遺伝的アルゴリズムを用いたルールの生成と削除を行う. しかし, 本論文における実験では存在しうるルールが多くないため, ルールの生成と削除は行わない.

2.2 グローバルなルール共有分類子システム

本論文ではグローバルにルール共有する分類子システムとして, Rule-Sharing Learning Classifier Systems (RS-LCS)を提案する. グローバルにルール共有する方

法は過去にも提案されており、これを提案すること自体は本論文の目的ではない。ここでRS-LCSを提案するのは、後述のローカルにルール共有する分類システムとの比較を容易にするためである。

RS-LCSはZCSのようにシングルエージェントのための学習システムではなく、各エージェントが一つずつ持つZCSを包括する学習システムである。ここで、RS-LCSのルール共有メカニズムはマルチエージェント間で共有するメモリを用いて実現するが、それをFig. 1によって述べる。Fig. 1は左から右へとプロセスが進む様子を表している。Fig. 1の左はすべてのエージェントのルール群を示しており、それぞれのルールは条件部・行動部・強度を持つことが示されている。

Fig. 1 Rule-sharing mechanism in RS-LCS

共有用ルール群を作成する集約のアルゴリズムは次のようになる。

- (1) 各エージェントについて
 - (i) 特定のルール数 (SN) をZCSのルーレット選択メカニズムに集約候補とする。
 - (ii) 重複するルールがあるときは重複分を集約候補から外し、重複がなくなるまで繰り返す。
- (2) 集約候補として (エージェント数 $\times SN$) 個のルールが挙げられる。
- (3) それらをFig. 1の中央の共有用ルール群に保持するが、下記の規則によって集約する。
 - 条件部と行動部のいずれも一致するルール (以下、単に一致するルール) はそれらすべてのルールの強度を平均し、一つのルールにする。
 - 学習分類システムの条件部には (ワイルドカードとしての) #が含まれることがあるが、この#もほかの記号と同様に扱う。すなわち、このルール一致の確認の時にはワイルドカードの働きはしない。

そして、集約された共有用ルール群は、Fig. 1の右に示すように各エージェントに反映される。そのアルゴリズムは以下になる。

各エージェントについて以下を実行する。

共有用ルール群のすべてのルールについて以下を実行する。

- その共有用ルール群のルールに一致するルールが、そのエージェントにある場合、

$$s \leftarrow s + \zeta(s_s - s) \quad (4)$$

のように強度を更新する。ただし、 s はエージェント側のルールの強度、 s_s は共有用ルール群側のルールの強度、そして ζ は反映率である。

- その共有用ルール群のルールに一致するルールがそのエージェントにない場合、条件部と行動部が同じで初期強度 S_0 を持つルールをそのエージェントに追加し、上と同じ処理を行う。ただし、このルールの追加に際してルールが最大数 (N) を越える場合は、ZCSの削除メカニズムにより削除する。そのとき、共有用ルール群によって更新されるルールは削除されない。

上記のルール共有メカニズムが実行されるのは、報酬がマルチエージェントに与えられたときである。しかし、報酬が与えられたときにルール共有メカニズムが必ず実行されるわけではない。頻繁なルール共有メカニズムの実行は、大きな計算コストを必要とするうえに、自律分散で活動するというマルチエージェントの前提にもそぐわない。また頻繁に実行しても、エージェントの学習は進んでおらず、あまり意味のないルール共有になってしまう。これらを考慮して、マルチエージェントに報酬が与えられた回数が、ある特定の回数 (Int_{global}) に達するごとにルール共有メカニズムを実行するとする。

2.3 ローカルなルール共有分類システム

本論文ではローカルなルール共有分類システムとして、Local Rule-Sharing Learning Classifier Systems (LRS-LCS) を提案する。このシステムは、各報酬ごとの任意の2体間の平均距離が得られることを前提条件としている。そして、平均距離が近かったエージェントは役割も近いであろうという見通しのもとに、LRS-LCSはそれらエージェント間でルールを共有する。

任意の2体間の平均距離は、マルチエージェントに対する各報酬ごとに行列の形で得る。具体的には、

$$M = (d_{ij}) \quad (5)$$

である。ただし、 d_{ij} は通し番号の振られたエージェント i とエージェント j の間の前回の報酬から当該報酬までの間の平均距離である。また $i > j$ についてのみ得られ、それ以外は0とする。

LRS-LCSはRS-LCS同様に報酬回数の一定のインターバル、 Int_{local} ごとにルールの共有を行う。RS-LCSではすべてのエージェントから共有ルール群が作られ、すべてのエージェントに反映されたが、LRS-LCSでは同じメカニズムで2体から共有ルール群が作られ、その2体に反映される。2体を選ぶアルゴリズムは以下である。

- (1) 平均の M を

$$\bar{M} = \frac{1}{Int_{local}} (\sum_{i=1}^{Int_{local}} M_i) \quad (6)$$

によって求める．ただし M_i は、最後にルールの共有が行われてから得られた報酬の回数が i のときの M である．

- (2) 以下を満たすようなエージェント i と j をルール共有のエージェントとして選ぶ．

$$d_{ij} = \min_{k,l} \bar{M}(k,l) \quad (7)$$

3. 単純化したサッカー

本論文では単純化したサッカーを実験に用いる．Fig. 2はその単純化したサッカーの概観を示している．フィールドは 30×30 の正方形のマスからなる．すなわち、離散空間である．左右二つのチームにエージェントは分けられ、ボールが配置される．実験によってエージェントの数は変わるが、左右のチームでその数は同じである．右端の縦のラインは左のチームにとってのゴールであり、ここをボールが越えると左のチームがゴールしたとみなされる．左端の縦のラインは右のチームのゴールである．

実験の開始時や、いずれかのチームがゴールしたあとには、ボールおよびエージェントは初期位置に置かれる．ボールの初期位置として、フィールド中央の四つのマスのいずれかがランダムに選択される．エージェントの初期位置として、左のチームのエージェントにはフィールド左半分内のマスがランダムに選択される．右のチームについては左右逆である．

Fig. 2 Simplified soccer

3.1 エージェントの行動

離散時間でエージェントは行動し、1単位時間ごとにエージェントは1マス動くか、あるいは動かない．その方向は上下左右であり、フィールドを越える場合は元に戻される．すべてのエージェントは同時に行動し、その各行動はZCSのルールの行動部を基に決定される．また、複数のエージェント、ボールは一つのマスに入ることができる．ボールとエージェントが同じマスに入ると、ボールは自動的に動き（蹴られ）、そのエージェントが属するチームのゴールに向かってボールが真横に動く．ボールと同じマスに2体以上のエージェントがいる場合

は、ランダムに1体が選択され、ボールが動かされる．ボールが動く距離は12マスであり、さらに上下左右2マス以内でランダムに動く．そのとき、ボールは上下のラインを越えない．

3.2 エージェントの知覚

実験においては左チームのエージェントのみを学習の対象にするため、以下は左チームのエージェントについてのみ述べる．

各エージェントの知覚はビット列からなり、その長さは7ビットである．この知覚のビット列が、ルールの条件部に利用される．

4ビットはチームメートの知覚にあてられる．各ビットは順に右、左、上、下が割り当てられ、その方向にチームメートがいると判断されたときに1がセットされる．上下左右の判断は以下ようになる．まず、エージェントを通る上下の直線上のマスとそれを左右に1マス広くした帯を考える．この帯上で、主となるエージェントより上にチームメートがいる場合は、上にいると判断する．逆の下の場合も同様である．つぎに、この帯より左と、帯上であってもエージェントの左隣のマスにチームメートがいる場合は左にいると判断する．逆の右の場合も同様である．チームメートが同じマスにいる場合は右にいると判断する．1がセットされなかったビットは0とする．後述の実験ではチームメートが見える距離という条件が与えられるため、チームメートのいる位置がその距離より小さくないと、そのチームメートは知覚されない．チームメートとの距離はマンハッタン距離である．

残りの3ビットはボールの知覚にあてられる．もしボールがエージェントの上下左右の直線上にある場合は、それぞれ011,100,111,000がセットされる．エージェントと同じマスにボールがある場合は000がセットされる．それ以外の場合、すなわちエージェントの右上、左上、右下、左下にボールがある場合はそれぞれ、010,110,001,101がセットされる．

より多くの知覚（より精緻なチームメートやボールの知覚、あるいは相手チームのエージェントの知覚など）を採用することで、より正確なエージェントの行動が得られることは予備実験で確認している．しかし、状態空間が大きくなることと、本論文の議論はこの7ビットで可能であることから、これ以上の知覚を採用しない．

3.3 エージェント間の平均距離

LRS-LCSが実装されたマルチエージェントは、報酬ごとに任意の2体のエージェント間の平均距離の行列 M を得る．具体的には、

$$M(i,j) = \sum_{t=1}^k \text{dist}_{t,ij} \quad (8)$$

である．ただし、 t は最後に報酬を得た後に経過した単位時間であり、報酬を得たのが k 単位時間後とする．また、 $\text{dist}_{t,ij}$ は経過時間 t における2体のエージェント i ,

j ($i > j$) 間のマンハッタン距離である。

4. 実験設定

4.1 右チームのエージェント

右チームのエージェントは貪欲な戦略をとる。すなわち、すべての行動でボールに近づく方向に進む。また、学習は行わない。これはすべての実験を通して常に同じである。一方で、左チームのエージェントは実験によって変わる。左チームのエージェントの比較は、この共通した右チームのエージェントに対する実験結果の比較により行う。

4.2 左チームのエージェント

左チームのエージェントに次の三つを用いる。

- (1) 各エージェントに ZCS を実装する。ルールの共有はしない。このマルチエージェントを以後 *ZCS* とよぶ。
- (2) RS-LCS を実装する。このマルチエージェントを以後 *RS* とよぶ。
- (3) LRS-LCS を実装する。このマルチエージェントを以後 *LRS* とよぶ。

ZCS のパラメータおよび Profit Sharing のパラメータ r は Table 1 のようにした。また RS-LCS および LRS-LCS のパラメータは Table 2 のようにした。RS-LCS のパラメータである Int_{global} は、実験によってエージェント数が異なっても一定である。一方で本論文の目的は、グローバル共有と比較したローカル共有の有効性の確認であるため、共有の頻度に関しては平等でなければならないと考える。このように考えると、ローカル共有を 1 回行うことで 2 体のエージェント間で共有が行われるため、全体が 2 体より多いときは頻度を変えないと平等ではないことになる。全体のエージェント数が 2 体のとき、ローカル共有では全体の 100% のエージェントが共有を行うので、グローバル共有と同じインターバルでよい。次に全体のエージェント数が 3 体のときは、ローカル共有で全体の約 66% のエージェントが共有を行うので、グローバル共有に 66% を掛けたより高い頻度で共有を行う。さらに、5 体のときは 40% である。

Table 1 Parameters for ZCS and r

Parameter	Value		
N	200	$P_{\#}$	0.33
S_0	0.0001	β	0.02
r	0.9	-	-

Table 2 Parameters for RS-LCS and LRS-LCS

Parameter	Value		
SN	100	Int_{local} (Agent 2)	100
ζ	0.5	Int_{local} (Agent 3)	66
Int_{global}	100	Int_{local} (Agent 5)	40

シングルエージェントと異なりマルチエージェントではエージェント間で報酬分配をしなければならない¹。以下でその方法について述べる。ここでは学習を行う左のチームのみを考える。まず単位時間ごとにボールを動かしたエージェントを記録する。ボールが動かなかったときは「なし」を記録する。そのうえで、チームがゴールした場合、最後にボールを動かしたエージェントに 1.0 の報酬が与えられる。さらに、ボールが動かされた記録を 1 単位時間さかのぼるごとに 1.0 に 0.9 を乗じていく。その中でエージェントがボールを動かした記録があるときは、そのときの大きさの報酬がそのエージェントに与えられる。一度与えられている場合は与えられない。すべての記録をさかのぼったら、記録を削除する。報酬が全く与えられなかったエージェントは 0.0 が与えられる。相手チーム（右のチーム）がゴールした場合は（左のチームの）エージェントに 0.0 の報酬が与えられる。それまでにボールを動かしたエージェントの記録も削除される。

4.3 実験の組合せ

実験の組合せを Table 3 に示す。Table 3 のように、左のチームの種類、チームメートが見える距離、エージェントの数のそれぞれを一つずつ組み合わせる場合の、すべての組合せについて実験する。ただし、チームメートが見える距離が 5 の実験はエージェント数が 3 のみであり、エージェント数が 2 の実験はチームメートが見える距離 0 のみである。

Table 3 Sets of experiments

Left Team	<i>ZCS</i> , <i>RS</i> , <i>LRS</i>
Sight for Teammate	0, (5), ∞
Number of Agents	(2), 3, 5

5. 実験結果

後に議論を行うために実験結果の提示を工夫する。より最適に近いルール群集合を左のチームのエージェントが得れば、より高い得失点差（（左のチームのゴール数）−（右のチームのゴール数））になることは 4 章の実験設定より明らかであるため、得失点差を左のチームのエージェントのルール群集合の評価に用いる。各実験は 3,000,000 単位時間を 10 回ずつ行う。そして、10,000 単位時間ごとに区切った得失点差を各実験の結果とする。

各実験の結果から、学習が収束する時点と収束する値を算出し、評価に用いる。その算出方法として本論文では以下を用いる。

¹ここでは単純化したサッカーに有効と思われる報酬分配法を述べるが、1 章で触れたように、マルチエージェントの報酬分配法は研究の途上であり、最善である保証はない。

収束時点：0 から 50,000 単位時間まで（前半）の得失点差の平均、および 50,000 単位時間から 100,000 単位時間まで（後半）の得失点差の平均をとる．（（後半の平均）－（前半の平均））／（前半の平均の絶対値）＜0.01 ならば、収束時点を 50,000 単位時間（前半と後半の境界）とする．もしこの条件が成り立たなければ前半と後半の範囲をそれぞれ正の方向に 10,000 単位時間ずつずらす．これを上記の条件が成り立つまで繰り返し、収束時点を得る．

収束値：各実験の最後の 500,000 単位時間の間の 10,000 単位時間ごとの得失点差の平均とする．

Fig. 3 は左チームの種類が ZCS、チームメートが見える距離 0、エージェント数 3 のときのある実験における得失点差のグラフである．横軸は単位時間数であり、縦

Fig. 3 A result example in Left team ZCS, Sight for teammate 0, Agent 3

Table 4 Result: Sight for teammate 0, Agent 3

Left team	Convergence Time		Convergence Value	
	Ave.	Dev.	Ave.	Dev.
ZCS	266,000	81,000	116.5	17.0
RS	86,000	20,000	91.2	35.7
LRS	190,000	137,000	124.7	33.0

Table 5 Result: Sight for teammate 5, Agent 3

Left team	Convergence Time		Convergence Value	
	Ave.	Dev.	Ave.	Dev.
ZCS	508,000	353,000	130.0	7.6
RS	285,000	78,000	136.7	17.8
LRS	329,000	182,000	137.2	8.5

Table 6 Result: Sight for teammate ∞, Agent 3

Left team	Convergence Time		Convergence Value	
	Ave.	Dev.	Ave.	Dev.
ZCS	612,000	190,000	138.4	2.4
RS	463,000	243,000	184.3	15.1
LRS	536,000	307,000	188.2	18.2

軸は各 10,000 単位時間ごとの得失点差である．Fig. 3 に収束時点の判定範囲、収束時点、収束値の判定範囲、および収束値の具体例が示されている．

以降では、10 回の実験における収束時点と収束値の平均および標準偏差を提示する．Table 4, Table 5, Table 6 はエージェント数 3 のときの、チームメートが見える距離 0, 5, ∞ の結果を示している．エージェント数 5 のときの結果は、エージェント数 3 のときの結果と傾向が同じであるため、付録 1 に示している．これらの表においては、それぞれの左チームの種類に対する収束時点と収束値の平均および標準偏差が示されている．

Table 7, Table 8 はエージェント数 3, チームメートが見える距離 0, ∞ において、LRS-LCS がルール共有したエージェント組合せの割合であり、典型的な例の一つである． $i \leq j$ は 0 であるが、“-”として省略してある．Table 9, Table 10 はエージェント数 5, チームメートが見える距離 0, ∞ の同様の結果である．

また、Table 11 はチームメートが見える距離 0, エー

Table 7 Example ratio of pair of rule-shared agents: Sight for teammate 0, Agent 3

Agent number	0	1	2
0	-	0.013	0.013
1	-	-	0.973
2	-	-	-

Table 8 Example ratio of pair of rule-shared agents: Sight for teammate ∞, Agent 3

Agent number	0	1	2
0	-	0.345	0.343
1	-	-	0.311
2	-	-	-

Table 9 Example ratio of pair of rule-shared agents: Sight for teammate 0, Agent 5

Agent number	0	1	2	3	4
0	-	0	0	0.433	0.001
1	-	-	0.151	0	0.261
2	-	-	-	0	0.153
3	-	-	-	-	0
4	-	-	-	-	-

Table 10 Example ratio of pair of rule-shared agents: Sight for teammate ∞, Agent 5

Agent number	0	1	2	3	4
0	-	0.153	0.154	0.081	0.071
1	-	-	0.112	0.089	0.098
2	-	-	-	0.071	0.095
3	-	-	-	-	0.075
4	-	-	-	-	-

Table 11 Result: Sight for teammate 0, Agent 2

Left team	Convergence Time		Convergence Value	
	Ave.	Dev.	Ave.	Dev.
ZCS	410,000	61,000	68.0	14.0
RS	115,000	30,000	55.2	22.7
LRS	115,000	30,000	55.2	22.7

エージェント数2のときの結果を示している。表の意味はTable 4等と同じである。

実際に獲得されたルール群集合の様子を示したものが、Table 12である。三つのタイプがあるのは、エージェントが獲得したルール群集合の性質により三つに分けられるという意味である。それぞれのタイプに特徴的なルールの内容が示されており、ボールが左方向にあるという条件に対する各行動（○は動かない）の正規化された相対的強度を表している。これらの値はある1回の実験結果である。Table 12のタイプ1は、すべてのエージェントにおいて、ボールが←にあるときは行動←が最も選ばれやすい、ということを示す。これは、均質なルール群集合を示している。タイプ2は、ボールが←にあるという条件のとき、行動→の強度が最も大きいルールをもつエージェントと、行動←の強度が最も大きいルールをもつエージェントがいることを示している。これは、非均質なルール群集合を示している。タイプ3は、すべてのエージェントにおいて、チームメイトが→にいて、かつボールが←にあるときは、行動←の強度が最も大きく、逆にチームメイトが←にいるときは、行動→の強度が最も大きい、というルールを持つことを示す。これは、均質なルール群集合を示している。

Table 13は、ZCS、RS、LRSそれぞれにおける10回の実験において、Table 12におけるどのタイプが得られたかを、チームメイトが見える距離ごとに示したものである。各エージェントが持つルールの強度は実験ごとに全く異なるが、Table 12に示したルールにおいて、最も強度の大きい行動が一致すれば同じタイプとした。

6. 議論

以下ではRS-LCSの結果を通してグローバルな政策共有システムの効果と弊害について議論する。そのうえでLRS-LCSの結果を通して、ローカルな政策共有システムがグローバルな政策共有システムの欠点を補っていることを示す。最後に、本論文で提案したローカルな政策共有システムの限界と今後の展開について議論する。

6.1 グローバルな政策共有システムの効果と弊害

本節以降の議論において、収束時点がより早い/遅い、あるいは収束値がより小さい/大きいと述べるときは、Table 4から6に示した単なる標本の平均値の比較ではなく、有意水準を0.05とした母平均の差の検定の結果に基づいている。また、ほぼ同じというときは検定では有

Table 12 Acquired rules

Type 1 (All agents have the same rule. (Homo.))

Condition		Strength for each Action				
Other Agent	Ball					
→ ← ↑ ↓	(←)	→	←	↑	↓	○
# # # #	1 1 1	0.16	0.38	0.16	0.16	0.14

Type 2 (Agents have the different rules. (Hetero.))

Condition		Strength for each Action				
Other Agent	Ball					
→ ← ↑ ↓	(←)	→	←	↑	↓	○
# # # #	1 1 1	0.44	0.14	0.13	0.12	0.16
# # # #	1 1 1	0.14	0.43	0.13	0.15	0.15

Type 3 (All agents have the same rules. (Homo.))

Condition		Strength for each Action				
Other Agent	Ball					
→ ← ↑ ↓	(←)	→	←	↑	↓	○
1 0 0 0	1 1 1	0.17	0.30	0.16	0.17	0.20
0 1 0 0	1 1 1	0.36	0.15	0.17	0.14	0.18

Table 13 Rule-group type

Left team	Sight for teammate								
	0			5			∞		
	Type			Type			Type		
ZCS	1	9	0	1	6	3	0	5	5
RS	10	0	0	0	0	10	0	0	10
LRS	1	9	0	0	6	4	0	0	10

意な差が得られなかったことを示す。

Table 4から6の結果より、収束時点はルールを共有しないZCSよりルールを共有するRSの方が早い。これはルールの共有によって探索対象の状態空間が絞られることを裏付けている。チームメイトが見える距離0のとき収束値はRSよりZCSが大きい。そして、距離5のときはほぼ同じである。一方で距離∞のときZCSよりRSの方が大きい。

上記の収束値における逆転を考えるにあたり、エージェントの協調行動の内容について考える必要がある。左のチームは、一度ボールを動かしただけではゴールに届かず、右のチームに戻されてしまう。したがって、左のチームは連続して少なくとも2回ボールを動かす、すなわちフィールドで右と左に分かれる必要がある。これは当然1体のエージェントでは達成できず、左のチームの2体以上が協調して行動しなければならない。もしチームメイトの位置が見えないならば、上記の協調行動を効率よく行うためにエージェントはフィールドの左右のどちらかを専門的に受け持つしかない。このようなルール群集合はTable 12のタイプ2となる。逆に、各エージェ

ントが持つルール群が同じであることをここでは均質とよぶ。上記の考察から、チームメートの位置が見えないときには、均質なマルチエージェントは効率よく協調行動ができない。このようなルール群集合はTable 12のタイプ1となる。

一方でチームメートの位置が見えるときには、フィールドの左右のうち、チームメートと比較して自分がより近い方に移動すればよい。これは各エージェントの持つルール群が左右のどちらかを専門とする必要がなく、同じでよいことを示す。これは均質なマルチエージェントも効率よく協調行動できることを示す。このようなルール群集合はTable 12のタイプ3となる。

Table 4のチームメートが見える距離0のとき、チームメートの位置が見えないので、ルール群集合は非均質、すなわちタイプ2の方がよい。実際にルール共有を行わないZCSは、Table 13において、タイプ1が1例、タイプ2が9例確認されている。一方で、RSは、Table 13において、タイプ1が10例である。上記は、ルール共有を行わないZCSが、ルール共有を行うRSより収束値において大きいという裏付けとなっている。

Table 6のチームメートが見える距離 ∞ のとき、チームメートの位置が見えるので、ルール共有を行わないZCSでも、ルール共有を行うRSでも、ゴールするための協調行動は可能である。すなわち、Table 13において、タイプ2も3も協調行動を実現できる。ただし、より効率のよいのはタイプ3である。なぜなら、動的に左右の役割を交代する均質のタイプ3に比べると、(左か右かの) 定位置に必ず移動するタイプ2はロスがあり劣るためである。ZCSは、Table 13において、タイプ2が5例、タイプ3が5例確認されている。これはZCSがタイプ2に由来する局所最適解に陥っていることを示している。一方で、RSはタイプ3が10例である。前述の通り、RSはタイプ2を得ることがない。そのため、RSはタイプ2に由来する局所最適解には陥らない。上記は、収束値においてはRSがZCSより大きいということを裏付ける。

Table 5のチームメートが見える距離5のとき、ZCSとRSでは収束値においてほぼ同じである。チームメートが見える距離5のときは、チームメートの位置が部分的に見えるので、ゴールするための協調行動は互いに近くにいるときのみ可能となる。これはタイプ2よりも3の方が効率が良いとはいえなくなる場合となる。ZCSは、Table 13において、タイプ1が1例、タイプ2が6例、タイプ3が3例確認されている。一方で、RSはタイプ3が10例である。前述の通り、RSにタイプ1が0例であることから、タイプ1よりもタイプ3の方が効率はよいとわかるが、タイプ2とタイプ3の効率はどちらがよいともいえない。上記は、ZCSとRSでは収束値においてほぼ同じということを裏付ける。

このように、RSすなわちグローバルな政策共有は効

果があるときと弊害を生じるときがある。このような効果と弊害の差が、本論文の単純化したサッカーで観察されたのは特殊なケースではなく、マルチエージェント学習におけるグローバルな政策共有では一般的に起こりうると考える。そして、このような効果・弊害の因果関係を事前に予測することは困難であり、グローバルな政策共有による弊害が生じることを事前に防ぐことは困難であるといえる。

本節の議論は3体の実験結果について行ったが、付録1に結果がある5体の実験、および予備実験で行った10体の実験でも同様の結果である。これは次節以降でも同じである。

6.2 ローカルな政策共有システムの有効性

Table 4の結果より、収束時点において、ローカルにルール共有するLRSが、グローバルにルール共有するRSよりは遅い、しかしルールを共有しないZCSよりは早い。また収束値においては、RSより大きく、ZCSとほぼ同じである。前述したように、チームメートが見える距離0のときは非均質なルール群集合である方がよい。ルール共有しているエージェントの組合せについて見てみると、Table 7の3体の例では、エージェント1と2においてのみルール共有を行うことで、ルール群集合が均質になることを避けつつルール共有している。また、5体ではTable 9からわかるように、この例ではエージェント0と3のグループ、エージェント1と2と4のグループでルール共有を行い、均質になることを避けている。この単純化されたサッカーにおいては、空間的に近いエージェントは報酬を得るための役割も近いため、そのエージェント間でルール共有を行うことは、役割の異なるエージェントがルール共有することを避けることになる。

このような共有を行っているため、Table 13において、LRSはタイプ1が1例、タイプ2が9例確認されている。これは、ZCSとほぼ同じ収束値であることを裏付けている。

チームメートが見える距離 ∞ のとき、Table 6の結果より、収束時点についてはTable 4同様の大小関係である。また収束値においては、ZCSより大きく、RSとほぼ同じである。前述したように、ZCSはタイプ2に由来する局所最適解に陥っており、それを避けたRSはZCSより高い収束値である。ルール共有しているエージェントの組合せについて見てみると、Table 8およびTable 10からわかるように、各エージェントの共有はほぼ平等に行われており、RSと同様に均質なルール群集合を持っていることになる。実際に、Table 13において、LRSはタイプ3が10例確認されている。これは上記の議論を裏付ける。

Table 5のチームメートが見える距離5のとき、収束時点についてはZCSよりも早く、RSとほぼ同じである。

そして、収束値については *ZCS* および *RS* とほぼ同じである。前述したようにこの条件においては、タイプ 2 よりも 3 の方が効率がよいとはいえない。Table 13 において、*LRS* は、タイプ 2 が 6 例、タイプ 3 が 4 例確認されている。上記は、*ZCS* と *LRS* では収束値においてほぼ同じということを裏付ける。

以上の議論から、本論文で提案したローカルな政策共有は、グローバルな政策共有で陥る局所最適解を避けて最適解により近い解を得られることが示せたと考える。

6.3 ローカルな政策共有システムの限界と今後の展開

LRS は、同じ役割を果たすであろうエージェントが冗長に存在することを前提としている。単純化されたサッカーにおいてはゴールに必要な役割は二つあるため、エージェント数 2 のときはそのような冗長なエージェントが存在せず、Table 11 に示すように、*RS* と全く同じになってしまう。このことから、同じ役割を果たすであろうエージェントが冗長に存在しないときには、本論文で提案するローカルな政策共有システムは、グローバルな政策共有システムの欠点を補えないといえる。

RS-LCS, *LRS-LCS* にはパラメータが存在するが、今回はその影響を十分には考察できていない。たとえば、ルール共有の頻度 (Int_{global} など) は 100 としていたが、より頻繁な 1 あるいは 10 を同じ実験環境において試した際にはほとんど差はなかった。しかしながら、1000 の場合はルール共有しない場合に近かった。このような関係性についてこれ以上の考察をしていないが、ルール共有にはコストがかかるので今後調べる必要がある。ほかに、*LRS-LCS* は報酬ごとの任意の 2 体間の平均距離が得られることに加えて、平均距離が近かったエージェントは役割も近いという条件が成り立つときに有効であった。今後はこのような条件が成り立たないときにも有効となるように改善を加える予定である。

7. 結論

本論文では空間内を自律的に移動するマルチエージェントの強化学習を取り扱った。そして、その状態空間の爆発への対策としてグローバルな政策共有を用いることの問題点を指摘した。その改善を目指し、ローカルな政策共有を行うシステムを学習分類システムに基づいて提案した。その検証には、離散空間・離散時間などの単純化を施したサッカーに基づく実験を用いた。そのうえで、(1) マルチエージェントが各報酬を得るまでの任意の 2 体間の平均距離が得られること、(2) 平均距離が近かったエージェントは役割も近いこと、(3) 役割に対して冗長なエージェントが存在すること、が成り立つときには、本論文で提案したローカルな政策共有がグローバルな政策共有の欠点を補うことを確認した。

本論文で提案したローカルな政策共有は、同じ役割を持つエージェントの集団に分けること、すなわち集団を

創発させることで効果をあげている。エージェント同士がフラットであることを前提にした学習ではなく、組織を構成していくことを前提にした学習を目指す際に、本論文は原初的な知見を与えると考える。

謝 辞

本研究は通信・放送機構の研究委託「人間情報コミュニケーションの研究開発」により実施したものである。また、本論文に多くの貴重なコメントを頂いた査読者に感謝の意を表す。

参 考 文 献

- [1] 荒井: マルチエージェント強化学習 - 実用化に向けての課題・理論・諸技術との融合-; 人工知能学会誌, Vol. 16, No. 4, pp. 476-481 (2001)
- [2] M. Tan: Multi-agent reinforcement learning: Independent vs. cooperative agent; *The 10th International Conference on Machine Learning*, pp. 330-337 (1993)
- [3] 高玉, 寺野, 下原, 堀, 中須賀: 組織学習に基づく分散分類システムを用いた創発的問題解決への接近; 計測自動制御学会論文集, Vol. 35, No. 11, pp. 1486-1495 (1999)
- [4] H. Holland: Properties of the bucket brigade algorithm; *The 1st International Conference on Genetic Algorithms*, pp. 1-7 (1985)
- [5] S. Wilson: ZCS: A zeroth level classifier system; *Evolutionary Computation*, Vol. 2, No. 1, pp. 1-18 (1994)
- [6] J. J. Grefenstette: Properties of the bucket brigade algorithm; *The 1st International Conference on Genetic Algorithms*, pp. 1-7 (1998)
- [7] 荒井, 宮崎, 小林: マルチエージェント強化学習の方法論 - Q-Learning と Profit Sharing による接近; 人工知能学会誌, Vol. 13, No. 5, pp. 609-618 (1998)
- [8] 宮崎, 山村, 小林: 強化学習における報酬割当の理論的考察; 人工知能学会誌, Vol. 9, No. 4, pp. 104-111 (1994)

付 録

付録 1. 参考資料としての実験結果

Table A1, Table A2 はエージェント数 5 のときの、チームメートが見える距離 0, ∞ の結果を示している。

Table A1 Result: Sight for teammate 0, Agent 5

Left team	Convergence Time		Convergence Value	
	Ave.	Dev.	Ave.	Dev.
<i>ZCS</i>	379,000	162,000	200.7	5.4
<i>RS</i>	76,000	35,000	171.8	43.3
<i>LRS</i>	342,000	258,000	204.4	29.4

Table A2 Result: Sight for teammate ∞ , Agent 5

Left team	Convergence Time		Convergence Value	
	Ave.	Dev.	Ave.	Dev.
<i>ZCS</i>	761,000	229,000	203.5	7.9
<i>RS</i>	328,000	178,000	220.3	13.1
<i>LRS</i>	497,000	210,000	224.5	5.4