2013年度　卒　業　論　文

Robocupサッカーにおける profit sharingによる行動観察学習の研究

指導教員：渡辺 大地 講師
三上 浩司 准教授

メディア学部 ゲームサイエンス プロジェクト
学籍番号　M0110395
堀越　惟志

2013 年度　　卒　業　論　文　概　要

論文題目

Robocup サッカーにおける profit sharing による行動観察学習の研究

メディア学部 **学籍番号：M0110395**	**氏名**	堀越　惟志	**指導教員**	渡辺 大地 講師 三上 浩司 准教授

キーワード	マルチエージェント, 強化学習, profit sharing 行動観察, 模倣, 価値システム

強化学習とは、エージェントが与えられた環境下において試行錯誤し、行動に対して与えられた報酬によって実行すべき行動を追及していくもので、機械学習の一種である。この強化学習による行為獲得には、獲得すべき行為が複雑であるほど膨大な探索処理や学習時間を要する問題が存在している。先行研究ではこの強化学習に、他者の行為を観察して対象の状態を推定し、それを自己の行動学習にフィードバックする方法を組み込み、この手法が安定して学習を発達させることを示している。しかし、観測した情報の利用方法として自己の学習に状態価値を用いており、膨大な状態数への対応に弱く、行動の学習にも時間が掛かるという問題がある。

本手法は、Robocup サッカー環境下において効率的な学習法の実現のため、profit sharing に他エージェントの行為を観察する方法を組み合わせた「行動観察 profit sharing」を提案し、これによりエージェントの行動学習が発達する事を示す。本手法の有効性を検証するため、RoboCup シミュレーションリーグの規定に基づいた Robocup サッカー環境のシミュレータに本手法を適用し、profit sharing と行動観察を組み合わせた学習が有効に機能することを示した。

目 次

図 目 次

表 目 次

第 1 章

はじめに

1.1　研究の背景と目的

コンピュータ上での学習については古くから研究がなされており、13 世紀には既にラモン・リュイによって、論理的に知識を生み出す「論理機械」が開発されている [1]。1956 年にはダートマス大学で開催されたダートマス会議にて、人間的知能をコンピュータに与える研究（人工知能研究）が学術分野として確立 [2] し、1959 年にはアーサー・サミュエルが機械学習を「明示的にプログラムしなくても学習する能力をコンピュータに与える研究分野」と定義した [3]。機械学習とは、全動作をプログラムせずとも、学習によって動作を獲得する力をコンピュータに与える、人工知能研究の一種のことである [4]。ダートマス会議に参加した多くの研究者は、人間と同レベルに知的な人工知能が近いうちに実現されると考えていたが実現出来ず、目標も曖昧だったことから 1973 年には研究への出資が取り止めとなった [5]。以降、人工知能研究は停滞と再興を繰り返し、現在まで続いている。

しかし、最近になって人工知能の学習機能が再び注目を集めている。その理由としては、コンピュータの性能向上が挙げられる。これによって、医療や機械対話、情報検索などの様々な現実の問題において、デジタル上でのよりリアルで複雑な環境構築が可能となった。その一方で環境の構築に掛かる労力も、環境の複雑さに比例して増加した。近年は将棋やオセロなどの人工知能は非常に洗練され、プロの人間に勝利することも不可能では無くなっている [6] が、これらの人工知能

は与えられる状況が限定的であり、環境が絶えず変化する現実の問題にはそのまま適用することが出来ない。そのため複雑な環境に対応する人工知能の手法として、機械学習の一種であり、ある環境下においてエージェントが試行錯誤し、結果として行った行動に対する報酬によって実行すべき行動を追及していくという方法を取る、強化学習 [7] が注目されるようになった。

強化学習を用いた人工知能の構築には、様々なシミュレーション方法が存在するが、その中でも本論文では、目的に対する手段が無数にあり、状況が膨大になり易いという点で現実問題に近しい、マルチエージェントを扱う Robocup サッカー環境下における学習法について研究する。

谷田 [8] は強化学習そのものは、特別にマルチエージェントのために提案されたものではなく、本来シングルエージェントに適した方法論であり、マルチエージェント環境に適用するためにはいくつかの問題点を解決しなければならないと述べている。その中の１つに、状態を目的状態に遷移する行動をしたエージェント以外へどれだけ報酬を分配するかという間接報酬問題がある。これはエージェント全体としての学習に大きな影響を与える問題であるが、宮崎ら [9] は直接報酬 R に対し、割引率 μ が非合理なルールの抑制範囲に当てはまるならば、間接報酬 μR が悪影響を与えない範囲で学習の向上に貢献することを示している。

また、マルチエージェント環境における学習手法の具体例として、高橋ら [10] の研究がある。高橋らはこの研究で、強化学習のに加えて、観察学習を用いている。観察学習とは、他者の行動やその結果をモデルとして観察し、自身の行動の参考とするものである [11]。この行為は、強化された、あるいはなにかしらの意思決定がされた結果として行動を起こした対象を観察するので、対象の状態を推測できれば、実質的には学習の試行回数が増加し、解の取得時間が短縮される可能性がある。この観察学習へ、行為一つ一つを別々に観察するモジュール構造型の学習法と状態価値を用いて、行為獲得と行為認識が行えることを示し、それによって学習の効率を高める手法を提案した。しかし、観察した行為を自己の行為に適用する場合、なんらかの適用基準を設けなければ、質の悪い行為で自己行為が上書きされ

てしまうことになる。これに関して高橋ら [12] は状態遷移系列における状態価値を比較し、確信度という値を導く方法を用いている。これによって、自らが判断した行為と観察した行為の優劣を比較することが可能である。だが、自身の学習方法として状態価値を利用する形態をとっているため、Robocup サッカー環境における状態数の設定や、学習速度に不安が残っている。一方で、Grefenstette[13] の研究では、環境の状態価値に依存せず行動を決定し、目的へ到達した時に報酬を与えることで高速に学習が行える、profit sharing という方法を提案している。この手法はエージェントの状態認識能力が不完全な状況下においても、学習が発達するという利点も持っているが、状態価値を利用しない特性上、学習を素早く行える代わりに、学習結果が局所最適解に陥りやすく、最適性が保障されないという問題がある。この局所最適問題を解決する研究として宮崎ら [14] の研究では、効率よく環境の状態を特定する ℓ-確実探査法と profit sharing を組み合わせた MarcoPolo という手法を提案している。しかし、この手法はマルチエージェント環境を考慮していない。

本研究では効率的に学習を行うための手法として、状態価値を考慮しない高速な学習を行える profit sharing へ、他者の行動を観察することで学習を効率化する方法を適用した、「行動観察 profit sharing」を提案する。検証の結果、RoboCup サッカーにおけるシミュレーションリーグ [15] の規定に基づいたサッカー環境をシミュレート出来る Robocup サッカーシミュレーションへ本手法を適用することで、既存の profit sharing と比較して、本手法は学習が効果的に行えることを示した。

1.2 本論文の構成

本論文の構成は全部で 5 章である。まず、第 2 章で既存の profit sharing の概要および問題点を述べ、第 3 章では Robocup サッカー環境と、提案手法となる行動観察 profit sharing の概要を解説する。第 4 章では実際に Robocup サッカーシミュレーションを用いて本手法の有効性を検証する。最後に、第 5 章で本論文の研究成果をまとめる。

第 2 章

profit sharingの概要

強化学習は、エージェントのある状況に対して正しい行動の基準を持たないため、教師なし学習に分類されている [16]。教師なし学習の目的は、状況データの背後に存在する本質的な環境構造を見つけ出すことであり、人間の想定外の優れた解の発見や、目的に対する手段を半自動的に生成することが可能である。その中で profit sharing(以下、PS) は、状態の価値を学習に利用しないために行動の網羅性に欠点があり、最適性が保障されないが、不確実性を持つ状態空間にも強く、高速な学習も可能である。本研究においては状態が多数存在する Robocup サッカーシミュレーション環境での学習を検証するため、この PS における高速性が必要となる。よって本章では、Grefenstette[13] の PS に関する具体的な解説を行う。

PS では、初期状態、または最後に報酬が与えられてから、次に報酬が与えられるまでの間に存在する、連続した状態行動対のルールをエピソードと呼ぶ。このエピソードの終了時、つまり目的状態への到達時に獲得した報酬を、エピソード内のすべてのルールに対して一括で割り当てる事で行動の優先度を変更し、学習を行っていく。割り当ての方法は、ルールに与えられた報酬を信用割り当て関数 f によって行動優先度の増分値に変換し、エピソード内で同ルールを複数回呼び出したならば、増分値を合算する。この合算した行動優先度の増分値を**強化値**と呼び、この強化値を計算するものを**強化関数**と呼ぶ。そしてエピソードの終了時、ルールに対して、元々の優先度に強化関数から得た強化値を加算することで行う。

状態が図 2.1 に示すような 4 × 3 マスの逆コの字型の迷路であり、最左上 S がエージェントの開始地点、S の 1 つ右を a、最左下 G をエージェントの目標地点、G のひとつ右を b、G のふたつ右を c とした時、エージェントの視界が周辺 8 マスであるとすると、b 地点の状況を a 地点として誤認してしまい、b 地点においても c 地点へ移動する行動を行ってしまう、いわゆるループ状態に陥る可能性が発生する。このような、状態をループさせる発端となるルールを無効ルールと呼び、そ

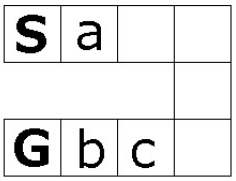

図 2.1: ループを誘発する迷路環境

れ以外のルールを有効ルールと呼ぶ。学習結果として有効ルールは無効ルールよりも f が大きい必要があり、次式 (2.1) を満たすならば、この条件が成り立ち、これを PS における合理性定理 [9] という。

$$L\sum_{j=i}^{W} f_j < f_{i-1} \forall\ i = \{1, 2, ..., W\} \tag{2.1}$$

ここで W はエピソードの長さを表す。この式 (2.1) は、目標状態に遷移する行動は有効ルールであり、このルールに与えられる f の値が、エピソード内の他ルールに与えられる f の値の合計に行動の候補数を掛けた値よりも多いならば、有効ルールはもっとも学習されているとするものである。この合理性定理に従う最も単純な f の決定法としては等比減少関数が知られており、これを用いるならば f の計算式は次式 (2.2) のようになる。

$$\begin{aligned} T &= W - 1 \\ f_t &= \gamma^{T-t-1} r_T \end{aligned} \tag{2.2}$$

ここで T は、エピソードの開始ステップを 0 とした時における終端ルールのステップ値、t は任意の現在ステップ値、γ は減衰係数である。宮崎らの合理性定理で無

効ルールが抑制される為には、γ の値が次式 (2.3) を満たす必要がある。

$$\gamma \leq \frac{1}{\max_{s \in S} |A(s)|} \tag{2.3}$$

等比減少関数による強化関数の設定は、目標状態に近くなければ、有効ルールごと無効ルールを抑制してしまう可能性のある方法であり、合理性定理を満たす γ の公式もまた、目標状態に隣接している状態における行動集合の数が多ければ、厳しく他のルールを抑制するものであるため、合理性定理を用いたアルゴリズムの学習態度は消極的であり、改善によってさらなる学習の効率化が行える可能性を秘めている。

第 3 章

提案手法

PS は、状態価値を学習のために必要としないので、状態数が膨大な環境下においても有効な手法である。しかし、局所解に陥り易いという問題点も内包しており、これを軽減した上で、より効率的な学習が行える方法を考察する。

3.1 Robocup サッカーシミュレーションの環境

Robocup サッカーシミュレーション (以下、RCSS) では横幅 105 メートル、縦幅 68 メートルからなるフィールド上でサッカーを行う。エージェントに現在サイクルにおける自己座標の状態を認識させるために、フィールドにおける座標は図 3.1 のような横 13 マス、縦 19 マスの四角格子配列へ変換する。ルールの裁定は RCSS を実行管理する RCSS サーバーによって行われ、おおよそ実際のサッカールールに準拠する。RCSS には最大 11 名のエージェントを登録し、サッカーをシミュレート出来るが、本研究では学習経過を顕著化するために登録人数を 6 名とした。時間の経過は 1 秒あたり 10 サイクルとし、各エージェントは 1 サイクル内で以下の行動を実行出来るとする。

- dash(dashPower):

  dashPower の分だけ現在方向へ自身を加速する。

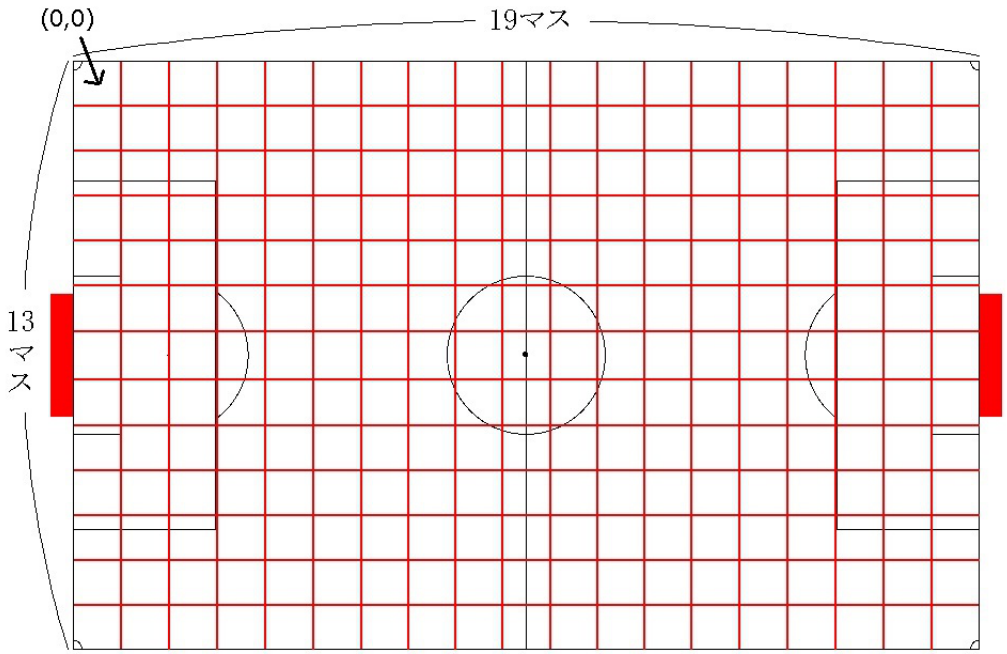

図 3.1: RCSS におけるサッカーフィールド

- kick(kickPower,kickDirection):

  現在方向に kickDirection を加算した方向へ kickPower の分だけボールを加速する。ボールが体に触れていなければ実行しても何も起きない。

- turn(turnDirection):

  現在方向に turnDirection を加算した方向へ自身が向き直る。

また、エージェントはスタミナを持ち、dash コマンドを実行するたび、dashPower 量に比例した分だけスタミナが減少する。スタミナはサイクル毎に少量回復するが、dashPower 量がスタミナを上回った場合、スタミナ以下の値まで dashPower は切りつめられる。

さらに、エージェントは視界を持ち、ボールへの方向と距離および、視界内にいる他エージェントの方向と距離、背番号を敵味方問わず受け取ることが出来る。この受け取る情報において方向を除くすべての情報は、対象との距離が遠い程に

ノイズが混じり、正確さが失われる特徴を持つ。これらに加えて本研究では、say および hear と呼ばれる特殊なコマンドを用いることで、遅延サイクル 1〜8 の範囲内でボール B の現在座標 $\rho(B)$ と味方プレイヤー F_i の現在座標 $\rho(F_i)$ を、自チーム内で共有・参照できるようにした。これにより対象を視界内に捉えずとも、ある程度味方やボールの位置が特定できる。

3.2 profit sharingによるRobocupサッカーシミュレーション環境のモデル化と学習

一般的に強化学習は、マルコフ決定過程（以下、MDP）によって環境をモデル化する [7] が、本手法においても一部それに従う。MDP とは、環境がとりえる状態の有限集合を $S = \{s_1, s_2, \cdots, s_n\}$ とし、エージェントが実行可能な行動の集合を $A = \{a_1, a_2, \cdots, a_n\}$ とする。この時ある行動 a を行った時、状態 s が状態 s' に遷移する確率を $Pr = \{s'|s_t = s, a_t = a\}$ または $P^a = (s, s')$ と表す。これに加えて、状態が遷移した時に与えられる報酬期待値 $R^a = (s, s')$ や、ある状況時にどんな行動を行えるかを表す $\pi(s, a)$ などを定義するモデルである。これをアルゴリズムによって解析することで、エージェントが学習する能力を得る。

本研究では、エージェント自身を M と定義する。また、部分マルコフ決定過程 [17] である RCSS の環境状態 s_t をマルコフ決定過程 [7] に近似させるために、状態を格子配列上の自己座標を表す 2 次元ベクトル $\rho(M)$、15 メートル以内に接近している敵の数 c、および自身がボールに最も近いかを表す論理型 b に分類し、これらを総じて e_t と表す。さらに、行動集合 A を以下のように設定する。

- pass():

  自分よりも前にいるエージェントの方向へボールを kick する。蹴る力はエージェント間の距離*4 とする。自分よりも前にエージェントが居なければ、最も近いエージェントへボールを kick する。

- dribble(dribbleDirection):

dribbleDirection 方向にボールを kick する。

- obtain():

  ボールの方向へ dash する。

- look():

  ボールの方向へ turn し、状況を観察する。

- move(t_x,t_y):

  $\mathbf{t} = \{t_x, t_y\}$ 方向へ turn し、$\rho(M)$ が $\mathbf{t}$ と等値になるまで dash する。

この時、A はそれぞれの行動が選択される確率を保持しており、R_p、R_d、R_o、R_l、R_m と表記する。e_t における b が true であるならば、$a_t = \{p = pass, d = dribble, o = obtain\}$ となり、b が false ならば $a_t = \{l = look, m = move\}$ となる。これによって a_t は表 3.1 の値を取りうるものと表すことが出来る。そしてこの a_t と e_t を、組み合わせた状態行動対を P_t とし、各エージェントのエピソード E_W に対して毎サイクル記録する。また直接報酬 R は、ボールを相手チームのゴールへ到達させたエージェントのみが獲得する。この R を獲得した場合、f_t を式 (2.2) に従って算出し E_W 内のすべての P_t に加算し、処理が終わった時に E_W を初期化 (W=0) し、再び記録を開始するようにした。

表 3.1: P_t の取りうる値

e_t				a_t
$\rho(M)_x$	$\rho(M)_y$	c	b	
0〜18	0〜12	1〜6	true	{p,d,o}
			false	{l,m}

3.3 a_t の計算方法

P_t には、毎サイクルごとに $\rho(M)$ と c、b の値を e_t として格納している。これらの値を用いて a_t に格納する行動を A の中から決定する。

b が false ならば、R_m と R_l を用いてルーレット選択を行い a_t を決定する。初期値は双方とも 1 とした。ここでルーレット選択とは、抽選する値すべての合算値を最大とした乱数を発生させ、それを元に行動を選択するものである [18]。b が true ならば、a_t には基本的に obtain を一意に格納するが、M がボールに接触した場合は dribble を一意に格納する。しかし、ボールに接触している上で $c \geqq 2$ であるならば R_d の初期値を 1 とし、R_p の初期値は以下の式 (3.1) によって算出し、R_d と R_p を用いたルーレット選択によって、a_t を決定する。また、dribble を a_t に格納した場合はさらに、蹴る方角を-60・-30・0・30・60 度の方角からルーレット選択する。詳細を表 3.2 へ示す。

$$R_p = \begin{cases} 2 & (3 \geqq c \geqq 2 \text{ のとき}) \\ 5 & (c \geqq 4 \text{ のとき}) \end{cases} \tag{3.1}$$

a_t を一意に決定する場合、例え報酬を得たとしても行動選択に影響が無く、これによる学習効率の低下を防ぐため、該当サイクルでは E_W に対して P_t は適用せず、エピソード長 W も変動しない。結果として E_W は次式 (3.2) のように計算することとなる。

$$\begin{aligned} W &= \begin{cases} W+1 & (a_t\text{が一意でないとき}) \\ W & (a_t\text{が一意のとき}) \end{cases} \\ E_W &= \begin{cases} P_t & (a_t\text{が一意でないとき}) \\ E_W & (a_t\text{が一意のとき}) \end{cases} \end{aligned} \tag{3.2}$$

表 3.2: a_t の決定法

b	ボールに接触	c	a_t
true	true	0〜1	d
		2〜6	d または p
	false	1〜6	o
false	true or false	1〜6	m または l

3.4 他エージェントの行動観察

高橋ら [12] の行動観察法は状態価値を用いるものであり、状態が不確実性を持つ環境および、状態価値を無視する PS とは相性が悪いと言えるため、状態価値を用いない行動観察法を定義した。

RCSS 環境下において他者行動を観察するために、以下の項目のような情報を本手法では設定した。

1. 観察対象
2. 自己行動への適用方法
3. 行動の推定方法

これらの詳細を節ごとに分けて以下に解説する。

3.4.1 観察対象

マルチエージェント環境においては、それぞれが独立して行動を起こしているため、全員が効果的に行動を実行している可能性は低い。特に RCSS においては目標状態への遷移にほとんど貢献しないエージェントが存在する可能性がある。そこで、目標状態に状態を近づけるために必要となる行動を報酬獲得への干渉行動と呼び、干渉行動を行ったエージェントを干渉エージェント I とする。本研究では干渉行動をボールへの接触として定義し、これを行った I を観察の対象とした。

3.4.2 自己行動への適用基準

各エージェントは、観察した I の情報を記録するために、E_W とは別に、観察エピソード $V_W = \{P_0, P_1, ..., P_{W-1}\}$ を保持する。他エージェントが干渉行動を行った場合に、そのエージェントを I として認識し、認識した時点から P_I を V へと記

録し始める。この時、V_W は以下の式 (3.3) に従うものとした。

$$
\begin{aligned}
W &= \begin{cases} W+1 & (I_t \neq M \quad \text{かつ } a_{I_t} \text{が一意でないとき}) \\ W & (I_t = M \text{ または } a_{I_t} \text{が一意のとき}) \end{cases} \\
V_W &= \begin{cases} P_{I_{t-1}} & (I_t \neq M \quad \text{かつ } a_{I_t} \text{が一意でないとき}) \\ V_W & (I_t = M \text{ または } a_{I_t} \text{が一意のとき}) \end{cases}
\end{aligned} \tag{3.3}
$$

このように、ボールに触れている $P_{I_{t-1}}$ のみを V_W へ記録することで、ゴールへ到達するための行動を効果的に取得することが出来る。すべてのエージェントが保持する V_W は、任意の干渉エージェント I_t が R を獲得した瞬間、I_t における E_W と同じく一斉に強化・適用し、これによってゴールへ到達するための行動を学習していく事となる。

各エージェントは独立して行動を行い、初期段階では行動基準が少ないため、等確率的に行動を選択しやすい。そのため、それぞれが学習の過程で局所解を得たとしても、その局所解同士が同一でないことが十分に有りうる。そのため、V_W を用いた学習は、学習速度が向上するだけでなく、単一的な局所解状態を緩和することも可能であると言える。

3.4.3 行動の推定方法

本研究では $\rho(B)$ と $\rho(F_i)$ を味方エージェント同士で共有・参照できる。よって、視界内にボールが存在しない場合はこの情報を参照し、視界内にボールが存在する場合は自身の視覚から得られる情報を参照した上で行動観察を行い、どのエージェントが I であり、なにをしているかを判断する。また、$\rho(B)$ および $\rho(F_i)$ は e_t を構成する状態に含まれないため、格子配列状として量子化をしていない、詳細な座標情報を格納した。

行動の推定は、1 サイクル前との状態比較によって行う。まずは $\rho(B)$ の位置に最も近いエージェントを距離の計算によって算出し、該当する $\rho(F_i)$ がボールに接触する範囲内に居るならば、I として認識する。次にボールの方向ベクトル $\nu(B)$ を 1 サイクル前のボール位置から計算し、さらに I から見たボールの方向ベクト

ル $\nu(B-I)$ を決定する。そして、この 2 つの値を元に、a_{t-1} は以下の式 (3.4) と (3.5) および (3.6) にしたがって決定する。

$$\mathbf{X} = \begin{cases} \frac{\nu(B-I)}{|\nu(B-I)|} & (|\nu(B-I)| \neq 0) \\ 0 & (|\nu(B-I)| = 0) \end{cases} \tag{3.4}$$

$$\begin{aligned} \mathbf{y} &= (\ \rho(B) + \nu(B)\ ) - (\ \rho(I) + \nu(B-I)\ ) \\ \mathbf{Y} &= \begin{cases} \frac{\mathbf{y}}{|\mathbf{y}|} & (|\mathbf{y}| \neq 0) \\ 0 & (|\mathbf{y}| = 0) \end{cases} \end{aligned} \tag{3.5}$$

$$a_{t-1} = \begin{cases} dribble & (\mathbf{X} = \mathbf{Y} \text{ かつ } I_t = I_{t-1}) \\ pass & (\mathbf{X} = \mathbf{Y} \text{ かつ } I_t \neq I_{t-1}) \\ obtain & (\mathbf{X} \neq \mathbf{Y}) \end{cases} \tag{3.6}$$

第 4 章

検証と考察

提案手法の有効性を検証するために、RoboCup シミュレーションリーグの規定に基づいたシミュレータに本手法を適用し、profit sharing と行動観察を組み合わせた学習が有効に機能するかの実験を行った。実験に関する詳細を述べると共に、実験結果に関する考察を行う。

4.1 実験概要

3000 サイクルを 1 ゲームとし、20 ゲーム分試合を行う事を実験 1 回分とし、学習アルゴリズムとして PS 法を採用したもので 1 回と、本手法を採用したもので 1 回の計 2 回の実験を行い、それぞれの実験においてどれだけの得点を獲得できるかの比較実験を行った。

自チームにおいては学習を顕著化するために、6 名の内前衛となる 3 名のアルゴリズムに対してのみ学習アルゴリズムを適応し、残りの後衛 3 名は、ボールが近くに来たらボールへ近寄り、接触したら 0 度の方角へ kick する単純なアルゴリズムを適応した。また、前衛 3 名は敵チーム陣のペナルティエリアへ接近した時、ゴールへボールをシュートするようにしている。敵チームのアルゴリズムは基本的に自チームの後衛アルゴリズムとほぼ同一であるが、kick する方向は自チームのゴールである点で異なる。開始時点での各エージェントの位置は図 4.1 のようになっている。

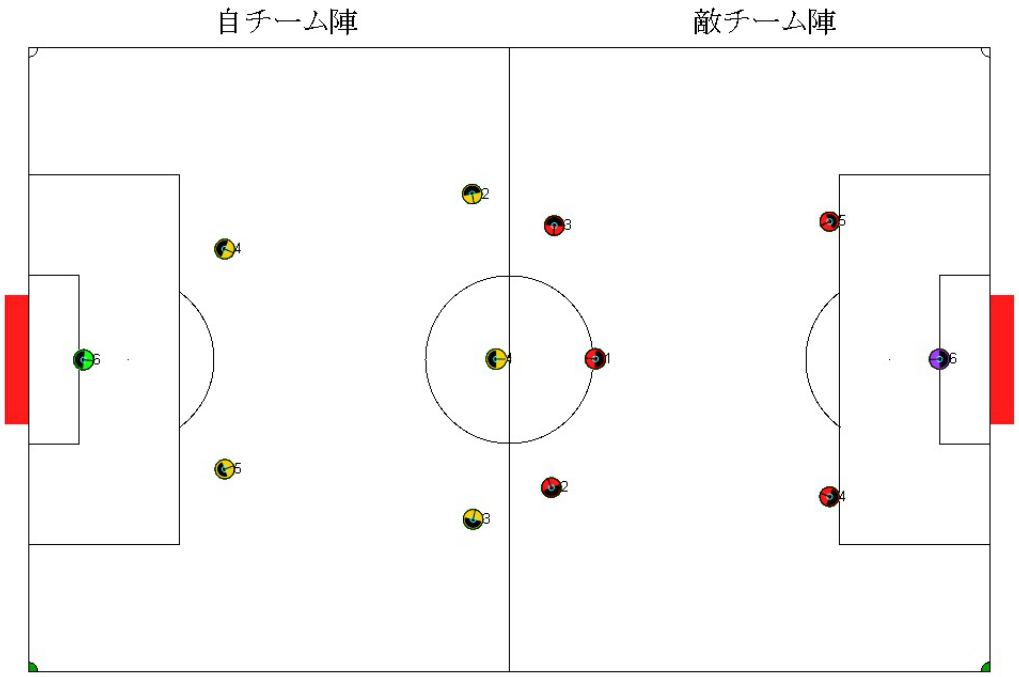

図 4.1: エージェントの基本位置

4.2 実験結果

実験結果として、学習に PS のみを適用した場合は、敵チームに対して平均得点数 6.6 対 6.8 という結果となり、行動観察 PS を適用した場合は、敵チームに対して平均得点数 5.8 対 4.4 という結果となった。PS のみの場合は自チームの得点数が徐々に上昇しているものの、敵チームの攻撃を抑えることが出来ず敗北した。これに対して行動観察 PS を適用した場合は、自チームの得点数が上昇しているのと同時に、敵チームの得点数を抑え、勝利している。各試合の得点数を図 4.2,4.3 に記す。

4.3 実験の考察

PS の特徴として、目標状態に近い状態ほど報酬を多く受け取るというものがあるため、学習過程を知るために、敵チーム陣におけるエージェントの行動を分析した。PS を適用した実験では、全試合中もっとも発生した同じ状況 (座標 (13,6)、

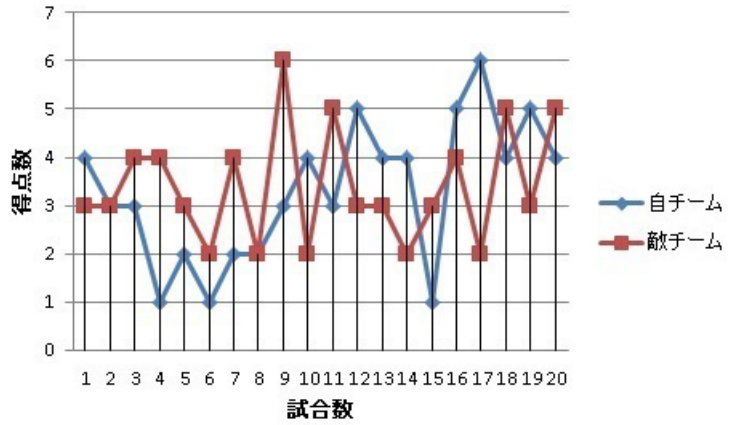

図 4.2: PS における得点の推移

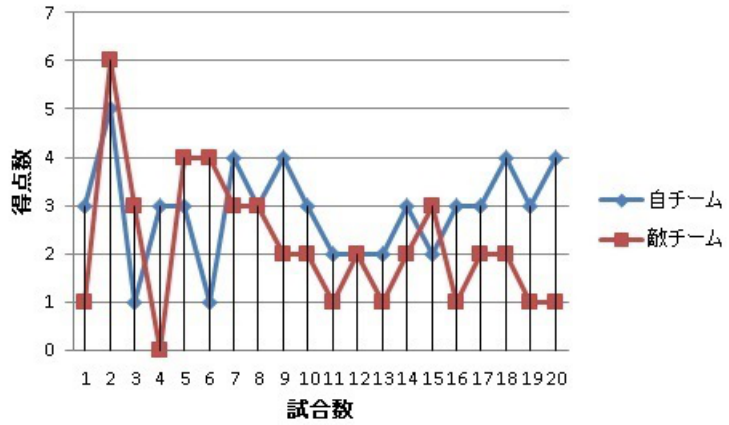

図 4.3: 行動観察 PS における得点の推移

敵数 2) において、4 試合目で 1 番のエージェントがボールを右上 30 度へ dribble する行動を取り、数サイクル後にゴールへボールを入れた。しかしその後、5 試合目で 2 番のエージェントが同じ状況になった時、1 番の行動を学習出来ていないため、敵ゴールキーパーめがけて dribble する行動をとってしまい、これによりボールが奪われて敵の得点を許してしまう事態が起きた。また味方の行動を学習しない性質のため、エージェントはそれぞれが統一性のない行動を取りやすく、ゴールへの到達ルートも図 4.4 に示すようにバラバラとなった。

一方で行動観察 PS は、PS のみの学習と同じようなゴール前での攻撃において、あるエージェントが右上へボールを逸らし、キーパーを避けてゴールした行動を

行った。そしてその後、他の味方エージェントが同様な状態におかれた場合も、同じ行動をとることが確認できた。またゴール前のみではなく、右サイドからの攻撃をする状態においても図 4.5 のように、ゴール出来た行動をエージェント同士が共有し、一定の攻撃ルートを確立していることが確認できた。

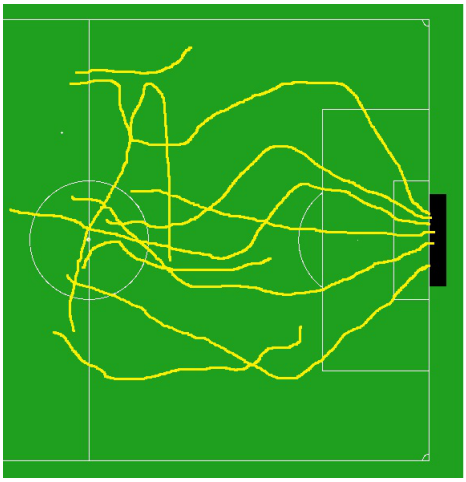

図 4.4: Profit sharing のみの攻撃ルート

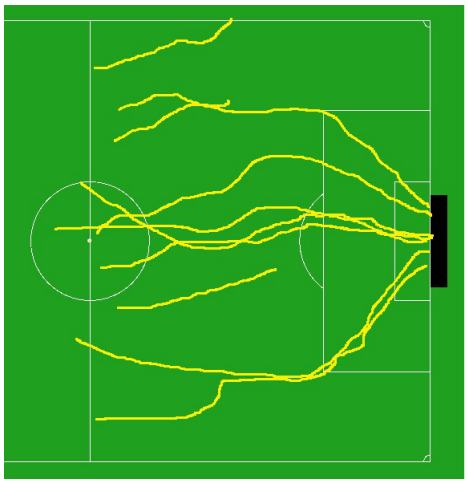

図 4.5: 本手法の攻撃ルート

このことから、本手法を RCSS 環境に適用することで、PS のみの手法と比べて有効に学習が行えたと言える。一方で、観察を行う対象となる行動はボールに接触しているものだけであるために、ボールを受け取るための立ち回りを学習することが出来ず、ゴールのための有効なパスが行われにくい事態が発生した。これに関しては、宮崎らの間接報酬 [19] などを参考にし、改善できると考えている。また全体を通して自チーム陣での動きの学習率が低く、後衛の kick に助けられている面があるが、これは報酬となるゴールが遠い目標であり、十分な分配報酬が得られないためと考えている。報酬の不足を解決するため、ゴール以外の目標をエージェントに与えることで学習率を上げる、中間報酬に関する研究 [20] なども精査する必要がある。

第 5 章

まとめ

本研究では、RCSS 環境下において効率的な学習法の実現のため、PS へ他エージェントの行為を観察する方法を組み合わせた行動観察 PS を提案し、提案手法の有効性を検証するため、RoboCup シミュレーションリーグの規定に基づいたシミュレータに本手法を適用した。実験の結果、PS だけの学習によるチームに比べて、本手法による学習を適用したチームがより効果的に行動することを示した。しかし、現状では pass 行動の方向が固定方向では無いために、まったく同じ行動の再現が難しく、適切ではない行動をとる問題や、dribble 中にボールを敵に奪われることを考慮していない問題、I でないエージェントとしての立ち回りは PS と変わらない問題などがある。今後の課題としては、土台となる環境のモデル化に使用した profit sharing 法における状況因子の特定法の改良や、エピソード報酬分配法の改良、観察学習のさらなる研究を行っていく所存である。

謝辞

本論文を作成するにあたって、多大なるご指導をして下さったゲームサイエンス・イノベーション研究室の渡辺大地先生と、三上浩司先生に心より感謝いたします。また、阿部雅樹先生をはじめとする研究室の先生・院生の方々にも深くお礼申し上げます。そして、共に励まし合いながら研究を進めてきた同研究室の友人達にも、感謝します。

参考文献

[1] Anthony Bonner. *The Art and Logic of Ramon Llull:A User's Guide*. BRILL, 2007.

[2] Daniel Creiver. *AI: The Tumultuous History of the Search for Artificial Intelligence*. Basic Books, 1993. pp.49-51.

[3] Alex Holehouse. Stanford machine learning. `http://www.holehouse.org/mlclass/01_02_Introduction_regression_analysis_and_gr.html`, 2011. (2014 年 1 月 5 日閲覧).

[4] Thomas Michell. *Machine Learning*. McGraw-Hill, 1997.

[5] Daniel Creiver. *AI: The Tumultuous History of the Search for Artificial Intelligence*. Basic Books, 1993. pp.100-144.

[6] 高橋 大介佐藤 佳州. 大規模な対局に基づいた教師データの重要度の学習. *GPWS*, Vol. 6, , 2012.

[7] 小林 重信木村 元. 強化学習システムの設計指針. 計測と制御, Vol. 38, No. 10, 1999.

[8] 谷田則幸. 人工市場とマルチエージェント強化学習. PhD thesis, 関西大学.

[9] 宮崎和光, 小林重信荒井 幸代. Profit sharing を用いたマルチエージェント強化学習における報酬配分の理論的考察. 人工知能学会誌, Vol. 14, No. 6, 1999.

[10] 高橋泰岳, 浅田 稔河又 輝泰. 自己の価値に基づく他者行為理解. 日本知能情報ファジィ学会誌, Vol. 21, No. 3, pp. 381–391, 2009.

[11] A. Bandura and R.W. Jeffery. Role of symbolic coding and rehearsal processes in observational learning. *Personality and Social Psychology*, Vol. 26, , 1973.

[12] 高橋泰岳, 浅田 稔田村 佳宏. 価値システムに基づく他者行為観察と自己行動学習の循環的発達. PhD thesis, 大阪大学大学院, 2009.

[13] J. J. Grefenstette. Credit assignment in rule discovery systems based on genetic algorithms. *Machine Learning*, Vol. 3, , 1988.

[14] 宮崎和光. 離散マルコフ決定過程における強化学習. PhD thesis, 東京工業大学, 1996.

[15] *Robocup Soccer Server Users Manual*, 2005.

[16] 伊藤友洋. 複雑ネットワークにおける経路学習問題に関する研究. PhD thesis, 滋賀大学.

[17] 小林 重信木村 元. 部分観測マルコフ決定過程下での強化学習:確率的傾斜法による接近. 人工知能学会誌, Vol. 11, No. 5, 1996.

[18] 辰巳 昭治河合 宏和. ルーレット選択を用いた Profit Sharing 強化学習における合理性についての一考察. PhD thesis, 大阪市立大学大学院.

[19] 宮崎和光, 小林重信木村 元. Profit sharing に基づく強化学習の理論と応用. 人工知能学会誌, Vol. 14, No. 5, 1999.

[20] 松井藤五郎. 自律型エージェントの行動学習に関する研究. PhD thesis, 名古屋工業大学, 2003.