サッカー型ゲームにおける 強化学習による犠牲的戦略獲得に関する研究

Sacrifice Strategy Acquisition by Reinforcement Learning in Soccer Type Game

國岡 将弘 * 亀井 且有 **
Masahiro Kunioka* Katsuari Kamei**

* ** 立命館大学 理工学部 情報学科
Department of Computer Science,Ritsumeikan University

Abstract: A purpose of our research is an acquisition of sacrifice actions in a multi-agent systems. The sacrifice action is not related to carrying out a task directly. However, it is an action helping other agents to complete their tasks. Moreover, communications between a sacrifice agent and other agents play an important role in the task completeness. In this paper, we verify that the sacrifice action is an important one in the multi-agent system and it is realized by the reinforcement learning through two experiments of soccer type game. First, we defines the sacrifice action and the task to be completed by multi-agent system in the experiments. Secondly, we show that the task is not alweys completed when there is no communication between agents. Thirdly, we show that the task is always completed by one agent's sacrifice action where there is commnication between agents. Finally, we discuss the detail behaviors of the sacrifice action and its usufulness in multi-agent systems.
キーワード：強化学習, マルチエージェントシステム
Keywords: reinforcement learning,Multi Agent Systems

1 はじめに

マルチエージェントシステムは, 複数のエージェントと呼ばれる行動主体が構成する系の総称である [1]. 今日, マルチエージェントシステムの分野では, 進化的手法や学習機構を実装したエージェントにより構成されたものが多く見られる．マルチエージェントシステムにタスクを与えた場合，エージェント間では共同作業や，タスク達成のための部分的な役割を担うといった相互作用が見られることがある．これらを協調行動と呼ぶ.

本研究では，協調行動の一つとして，直接タスクの達成とは関係の無い補助的な行動を取り上げる．これはあるエージェントがこの行動を取ることにより，他エージェントによるタスクの達成が可能になり，結果的にシステム全体として総合的なタスクの達成に至る行動である [3]．従来のマルチエージェントシステムの研究では，このような行動についての検証が少ない．このため，この補助的な行動を犠牲行動と定義し，有効性について検証することが本研究の目的である．

一方，協調について考えた場合，互いの状況を理解できないシステムでの協調行動は，学習過程で偶然獲得されたものであり，意図的でない．また，与えられたタスクの達成において，最適であるかの判断は困難である．そこで，本研究はエージェント間で情報伝達を行い，互いの状況を理解し行動を決定する．さらに，情報伝達による学習および獲得する行動への影響の検証を行う.

本稿では，サッカー型ゲーム環境を設定し，シミュレーション実験を通して犠牲行動の獲得の検証を行う．また，情報伝達の有無に関して獲得行動の比較により，情報伝達の有効性を検証する.

2 問題設定

2.1 実験環境

実験環境として図 1 のような環境を設定する．2 体の FW エージェント (以下 FW) は Goal に到達することを目的とし，GK エージェント (以下 GK) は,FW の侵入に対し防御行動をとる．実験中の 1 ステップとして，FW1 → FW2 → GK の順番で，それぞれ上下左右のいずれかに 1 マス移動か停止を選択し，行動する．

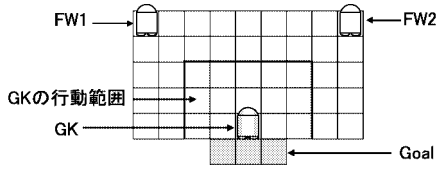

図 1: 実験環境

1 試行の終了条件を以下に示す.

- FW の一台が Goal に入った場合
- FW の一台が GK の 4 近傍に入った場合
- 1 試行の最大ステップ数を超えた場合

FW は自分を中心とした 5 × 5 の範囲が視界である．学習方法は Q-Learning を用い，報酬は FW の試行中における全ての行動に対して与えられる.

GK は図 1 に示す行動範囲内のみ移動可能であり，自分を中心とする 7 × 7 の視界に存在する FW を追跡目標とする．FW が視界に存在しないときは停止を選択する．また，FW が同時に 2 体存在するとき，以下に示す優先順位に従い，追跡する FW を決定する． また，優先順位の項目に該当しない場合，停止を選択する.

1. GK に近い FW を選択
2. Goal に近い FW を選択

追跡する FW を決定後，どの方向へ移動するかを決定する．図 2 に示すように，GK の視界を 4 つの領域に分割する．GK は追跡する FW が属する領域へ移動するものとし，各領域の①〜④の方向へ移動する．ただし，移動後の位置が GK の行動範囲外のときは停止を選択する.

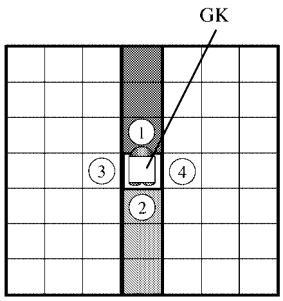

図 2: GK の視界

3 犠牲行動の定義

マルチエージェントシステムにおけるタスクの達成を考えるとき，エージェント間の協調行動の一つに，タスクの分業化がある [2]．従来のタスクの分業化に関する研究では，タスクに直接関係した行動を取るエージェントで構成されることが多い．本研究では，それぞれタスクに関係した分業化ではなく，タスクを直接達成する行動と，その行動に対する補助的な行動を行うといった分業化を考える．このとき，この補助的な行動を取ったエージェントは直接タスクを達成することができないが，システム全体の総合的なタスクの達成に至ることが期待される．そこで，この行動を本研究は犠牲行動と定義する．

ここで，本研究で設定する実験環境での犠牲行動の一例を図 3 に示す．このときの FW1 の行動は，FW2 が Goal に到達するために，GK を引き連れているため犠牲行動となる．

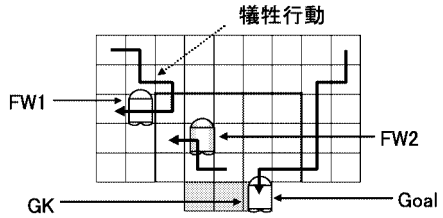

図 3: 犠牲行動の一例

4 学習法

エージェントの学習には様々な手法があるが，中でも広く研究に用いられている手法に強化学習がある．強化学習では式 (1) に示される時刻 t における強化信号の大きさを r_t とするとき，エージェントは現在から未来にわたる強化信号の重み和を最大化するように行動する．

$$V_t = \sum_{i=t}^{\infty} \gamma^{i-t} r_i \qquad (1)$$

ただし，γ は $(0<\gamma<1)$ なる定数であり, 報酬の割引率と呼ばれる．

Q-Learning[4] は, 状態と行動の組に対する評価を見積もる．この評価を Q 値と呼び, 状態と行動の組から評価の見積もりを導く関数を Q 関数と呼ぶ. 時刻 t における状態 x_t にあって, 行動 a_t を選択した結果, 状態は x_{t+1} となり, 強化信号 r_t が得られたとすると, 更新される Q 値の変更幅は次の式 (2) で定義される.

$$\Delta Q(x_t, a_t) = \alpha(r_t + \gamma max_b Q(x_{t+1}, b) - Q(x_t, a_t)) \quad (2)$$

α は学習率であり,$0<\alpha\leq 1$ なる定数である. 式 (2) は, 次の状態で最適と思われる行動を選択した時に得られる評価の見積もり $max_b Q(x_{t+1}, b)$ を一段階だけ割り引いた値と直接得られた r_t を加算したものに $Q(x_t,a_t)$ を近づけるという意味である．これらにより $max_b Q(x_t, b)$ を最適な行動を取った時の V_t に近づけることができる [5].

しかし，エージェントの学習法に多く見られるタスクを達成したエージェントのみが報酬を与える場合, 図 3 に示す犠牲行動を取ったエージェントが，報酬を受けることができない．この結果，エージェントは直接タスクを達成する行動のみ学習するため，犠牲行動を獲得できない．しかし，システム全体で見ると，総合的にタスク達成に貢献していると考えられ，この行動を学習させる必要がある．

そこで，本研究では報酬を一方の FW が Goal に到達した場合，もう一方の FW にも同様に与えることとする．この手法により，タスクを直接達成していない犠牲行動にも，効果がある行動として捉えることができる．

5 シミュレーション実験

5.1 実験 1

図 1 に示した環境を用いて実験を行った．図 4 に FW1 と FW2 における 1000 試行毎の Goal 到達率を示す．同図より，学習終了時における FW1 と FW2 の Goal 到達率に差が見られない．また図 5 に示す，1000 試行毎の Goal 到達に要した平均ステップ数を示す．図 4 および図 5 より Goal 到達率と平均ステップ数の増減関係について以下に詳しく検討する．

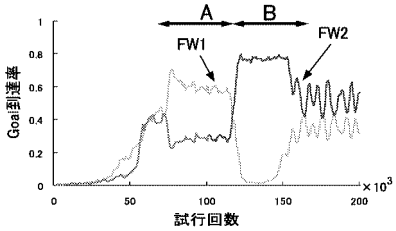

図 4: 1000 試行毎の Goal 到達率

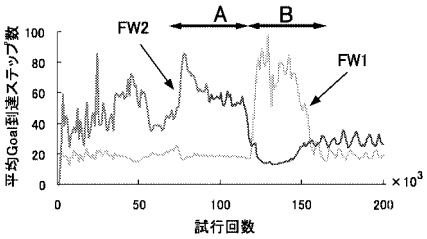

図 5: 1000 試行毎の平均 Goal 到達ステップ数

5.2 実験 1 の検討

図 4 より，実験初期における FW1 の Goal 到達率が，FW2 に比べ多いことがわかる．この時，図 6 に示す状況が観測された．まず，FW2 の視界は GK より狭く，一度 GK の視界に入ると FW2 は追跡され続ける．その結果，FW2 は GK から逃げる行動が多く学習され，優先的に行動を取り始める．一方，FW1 は直接 Goal へ向かう行動を学習することになり，両者の Goal 到達率に差が生じたと考えられる．

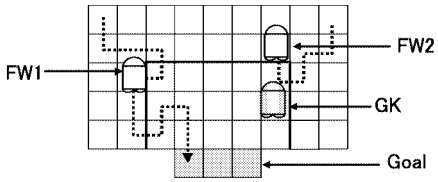

図 6: 期間 A

次に，図 4 の期間 A から B への Goal 到達率の逆転について考察する．この期間において図 7 に示す状況が観測された．FW1 は Goal に到達する最適な行動を学習するため，FW2 よりも早期に GK の視界に入ってしまい追跡される．しかし，FW1 は，GK が視界に存在するときの行動を学習していないため，ランダム探索として行動を模索する．さらに，GK が視界に存在する行動として，新たに GK に捕まらない行動を学習をする．その結果，Goal に向かうことが困難になり，Goal 到達率が減少する．また，FW2 は GK の追跡が無く，新たに行動を模索することで Goal に到達し，その後の学習から Goal 到達率が増加する．また図 5 より，到達率の逆転直後から Goal に到達する FW2 の到達ステップは減少し，追跡される FW1 は増加している．

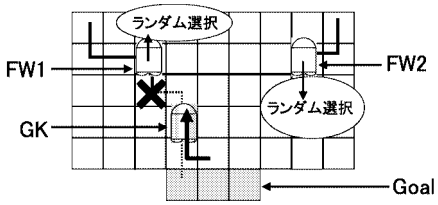

図 7: 期間 A から B の状況

図 4 の期間 B 以降の各 FW は，Goal に向かって行動し，GK に追跡されると逃げる行動を取る．このため，犠牲行動を確実に行う FW が存在しないことから，GK の視界に入る確率が等しくなる．そして，GK に追跡されない FW が Goal に到達するため，到達率に差が生じない結果になる．この原因として，各 FW が GK の存在の有無という情報でしか行動を選択できないことから，他方の FW の状態を考えた行動を取ることができなかったことが考えられる．したがって，各 FW は，自分の情報だけではなく，他方の FW における GK の情報を知るため，それぞれが情報を伝達し合う必要がある．

5.3 情報伝達

エージェント間に情報を伝達することを考える．エージェントが互いの状況を理解できない環境において獲得する協調行動は, タスクの達成に対して，最適な行動である保証はない. この理由として，各エージェントが自分の情報のみで行動する環境での協調行動は，偶然獲得されたものであり，意図的ではなく，環境全体の利益を優先した行動ではないことが挙げられる．このため，互いの状況を理解し行動することが必要であると考えられる [6].

そこで，情報伝達の導入を提案する. 情報伝達とはエージェント間で，現在における自分の状況を他のエージェントへ伝達することである. 情報を受け取るエージェントは, 情報を発したエージェントの状況を理解して行動を学習することができる．この情報伝達の導入により，エージェント間の相互理解から，現在の状況における最適な行動の選択が期待できる．

5.4 実験 2

実験 1 と同じ環境に加え，情報伝達を各 FW 間に導入し，実験を行った．この実験における情報伝達は，FW 間で互いの視界における GK の存在を伝達する．情報を受け取った FW は，自分の視界における状態と，他方の FW からの情報により，行動を決定する．図 8 に，FW1 と FW2 の Goal 到達率を示す．同図より，学習終了時の Goal 到達が明確に別離していることがわかる．ここで，1000 試行毎の Goal 到達に要した平均ステップ数を図 9 に示す．この図より，Goal に到達する FW2 のステップ数が収束していることがわかり，常に Goal に到達する最適な行動を獲得したと考えられる．

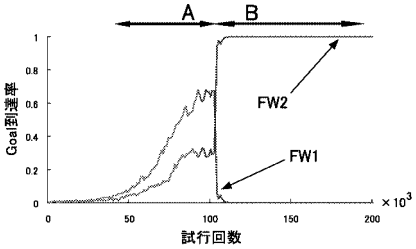

図 8: 1000 試行毎の Goal 到達率

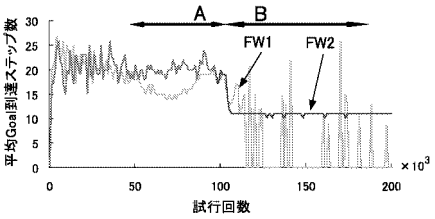

図 9: 1000 試行毎の平均 Goal 到達ステップ数

実験結果から，情報伝達による行動の影響について考察する．

まず，図 8 における期間 A から B の Goal 到達率の逆転については，情報伝達の有無に関係無く生じており，各 FW の行動も実験 1 と同様のことが考えられる．

つぎに図 7 に示す期間 A から B へのランダム探索以降では，FW1 は，常に追跡される行動を取り，かつ GK に捕まらない行動を獲得している．一方，FW2 は GK に追跡されない行動を取ることで，Goal に到達する行動を獲得している．したがって，FW の行動が別離した理由として，各 FW がそれぞれ異なった GK の情報を受け続けながら学習を進めたことが挙げられる．このとき，FW1 は犠牲行動として，GK が視界に存在し続ける行動を取り，また反対に FW2 は，GK が視界に存在しないため Goal に到達する行動を取る．

以上から，情報伝達を行うことで，それぞれの状況に対して最適な行動の選択が可能になり，その結果，FW1 の明確な犠牲行動の獲得が確認された．

実験において各 FW が獲得した Goal までの過程を図 10 に示す．試行開始後，各 FW は，図 10 の A 地点に移動後，停止を続ける．GK は FW1 を追跡目標とし，B 地点まで移動する．GK が B 地点に移動した次のステップにおいて，FW1 の視界には GK が自分の視界に確認される．そこで，FW1 は FW2 に GK が存在する情報を伝達する．次に FW1 は GK を引きつける行動として上方向に移動し，停止し続ける．一方，FW2 は情報を受け取ると，Goal に向かう行動を取り，最短ステップで到達する．

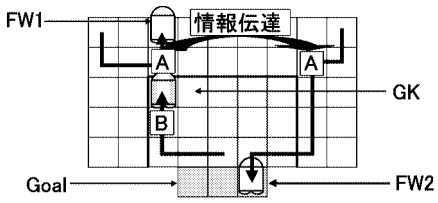

図 10: 情報伝達を行ったときの Goal 到達過程

6 おわりに

本稿では，マルチエージェントシステムにおける協調行動の一つとして，犠牲行動を取り上げ，その有効性を述べた．また，協調行動の獲得に関する問題点である，エージェント間の相互理解の必要性から，情報伝達を提案した．犠牲行動の獲得実験として，サッカー型ゲーム環境を設定し，実験を行った結果，情報伝達を行うことで，より明確な犠牲行動の獲得ができることがわかった．獲得過程における情報伝達の影響についても詳細に考察した．

情報伝達の導入により，状態数が増加する問題が生じる．強化学習において，状態数の増加は学習速度の低下などの理由から問題とされている．このため，状態数が増加しない情報伝達が必要となる．今後の課題として，情報伝達の問題を解決する手法を提案し，より巨大な環境で実験を行うことが挙げられる．

参考文献

[1] 嘉数 侑昇, マルチエージェントシステムの研究動向，システム制御情報学会誌, Vol.41,No.8,pp.291-296,1997.

[2] 畝見達夫, 強化学習エージェントの集団行動，マルチエージェントと協調計算III，日本ソフトウェア科学会 MACC'93, 近江科学社，pp.137-150,199.

[3] 白川 英隆，木村 元，小林 重信，強化学習による協調行動の創発に関する実験的考察，知能システムシンポジウム，Vol.25,No.6,pp.119-124,1998.

[4] C. Watkins, Technical Note Q-Learning, Machine Learning,Vol.8,pp.279-292,1992.

[5] 畝見達夫，強化学習. 人工知能学会誌，Vol.9,No.6,pp.830-836,1994.

[6] 河田 洋平，大倉 和博，上田 完次，自律エージェント行動戦略の進化的獲得における通信を影響に関する研究，システム制御情報学会研究発表会，Vol.44,pp181-82,2000.

論文受付番号 i018

問い合わせ先 〒 525-8577
滋賀県草津市野路東 1-1-1
立命館大学大学院理工学研究科
國岡 将弘
TEL:077-561-2807 FAX:077-561-2669
E-mail:kunioka@spice.cs.ritsumei.ac.jp