琉球大学学術リポジトリ

強化学習を用いた共同注視点に基づく合意形成

メタデータ	言語: 出版者: 琉球大学工学部 公開日: 2010-01-13 キーワード (Ja): キーワード (En): Focal-Point, Reinforcement leaning, Profit Sharing, Soccer Server 作成者: 亀島, 力, 与那覇, 賢, 遠藤, 聡志, 山田, 孝治, Kameshima, Chikara, Yonaha, Satoru, Endo, Satoshi, Yamada, Koji メールアドレス: 所属:
URL	http://hdl.handle.net/20.500.12000/14729

強化学習を用いた共同注視点に基づく合意形成

亀島 力*，与那覇 賢*，遠藤 聡志*，山田 孝治*

A consensus making method using focal points with reinforcement learning

Chikara KAMESHIMA*, Satoru YONAHA*, Satoshi ENDO*, Koji YAMADA*

Abstract

"Focal Point" is a method of sharing information without explicit communication between cooperative multi-agents. It is hard for an artificial agent to prescribe the fixed algorithm which decides a focal point without relation to the environment. In this paper, we suppose the focal points as a distribution of a situation to induce the behavior of the agent. The autonomous agent can get focal points as a learning of the inside model of itself by the feedback from the environment. Consequently, the mutual agreement of multi-agent is emerged from an unity of the probabilistic choice of fields. Each agent is reinforced the filed potential value respectively by successful task. Finally, focal points can be acquired as potential distribution map of reinforcement value.

Key Word: Focal-Point, Reinforcement leaning, Profit Sharing, Soccer Server.

1.はじめに

複数のエージェントによってタスクを処理するマルチエージェントシステムにおいて、エージェントは協調的、組織的に行動する必要があり、エージェントらの合意形成による最適な意志決定が必要である。しかし、環境情報に不確実性が含まれている場合や、明示的な通信が困難な環境では、合意形成による協調作業の実現は容易ではない。そのような場合、状況や場の必然からエージェントらの選択が自然に一致する点である共同注視点の概念を利用した合意形成が有効である [1]。

Schellingは社会的意志決定の場において、エージェント間の相互作用によってエージェントが特に興味を示す顕著な解をフォーカルポイントと定義した。我々はこれを、社会的環境を共有する複数の主体の間で選択が一致することから、共同注視点ととらえた。本研究では、合意形成を確率的な選択の一致として扱い、Profit Shaing[4]を用いた強化学習を基に、問題に対して適切な共同注視点を獲得する手法を提案する。

また、合意形成問題としてサッカーにおけるパス協調問題を取り上げる。パス協調問題では、プレイヤ間でパスコースに関する合意形成が重要になる。敵ゴールへのシュートを成功させるためには、敵味方の位置関係によって味方プレイヤへの直接パスが有効な場合や、敵プレイヤのいない場所へのパスが有効な場合がある。パスを出す場所を共同注視点とした合意形成を行うエージェントを設計し、計算機実験によって提案手法の有効性を検証する。

2.共同注視点

複数のエージェントが、処理対象の中のいずれに注目すべきかをその場の環境に応じて選択することを考える。このような状況において適応的な協調行動を実現するためには、各々のエージェントの選択対象の結果をできるだけ一致させたいという要求がある。この複数エージェントの共同注視点は、フォーカルポイント(Focal Point)あるいは焦点と呼ばれ、社会的な場の状況、構図などから規定されるものとして、さまざまな社会的局面に調査研究が行われている。

Schellingは、[3]において、さまざまな場面の社会的意思決定の事例調査の結果、表面的には同等な多くの解候補の中から、多くの人が場の状況や文脈情報に依存して、特定の解に引き込まれる傾向があることを示した。以下にその例を示す。

受理：1999年6月7日
* 工学部情報工学科
(Dept. of Information Engineering, Fac. of Eng.)
** 大学院理工学研究科情報工学専攻（1998年・修了）
(Graduate Student, Information Engineering)

Schellingの例題：１００枚の硬貨の分割
二人の被験者にそれぞれ１００枚ずつの硬貨を与えて互いに隔離し、「自分の硬貨を任意の枚数ずつ２つの山に分けよ」と指示する。このとき、被験者の関心事は自分の分割方法が相手の分割方法と一致するかどうかである。なぜなら、２人の分割が一致すれば２人にはご褒美が与えられ、一致しなければ何も与えられないからである。

Schellingは、上記の情況が与えられると、ほとんどの被験者は５０枚ずつの２つのグループに分割することを発見した(図１)。その理由として、被験者が客観的に分割の方法を考えた場合、どのような分割の仕方も同等であると考えられるため、何か唯一性を持つような分割（ここでは、対称性を満たす分割）を見つけ、相手も同様に考えるであろうと期待したことが考えられる。

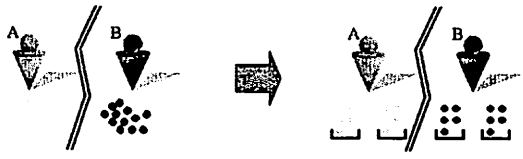


図１：硬貨の分類問題

ゲーム理論のような形式的な手法を用いても、なぜ人々は多くの同等の解の中から、ある特定の解を好むのかを説明することは困難である。Schellingの例題に対して均衡点の概念を導入すると、利得を得るためには被験者A,Bの分割が一致しなければならないゆえ、均衡点は

$$(A_{(0,100)}, B_{(0,100)}), (A_{(1,99)}, B_{(1,99)}), \cdots, (A_{(1,99)}, B_{(1,99)}), (A_{(100,0)}, B_{(100,0)}) \tag{2.1}$$

ただし、被験者Aの分割の戦略を$A_{(m,n)}$で表す

であり、その数は101個となる。これらの101個の均衡点は、被験者の利得の観点から見ても同等であり、どの分割を選ぶべきかを決定することができない。しかし、人間は巧みに相互作用の文脈情報を用いて、効率よく整合することが可能であることから、人間は現実社会における生活を通じてフォーカルポイントを決定する仕組みを既に獲得していると考えることができる。

3.合意形成の定式化

合意を形成する主体が人工エージェントである場合、フォーカルポイントを決定する固定的なアルゴリズムを、環境とのかかわりを抜きにしたまま、事前に規定することは困難である。そのため、エージェントの置かれている環境内での観測を通じてフォーカルポイントを決定する仕組みを自律的に獲得する枠組が重要となる。本研究では、そのようなフォーカルポイントの自律的な獲得のための手法を提案するために、合意形成のメカニズムを確率的な選択の一致という観点から定式化する。

いま、エージェントが取りうる選択の集合$\boldsymbol{A}$を

$$\boldsymbol{A} = \{a_1, a_2, \cdots\} \tag{3.1}$$

とする。エージェントＡとエージェントＢがそれぞれ選択a_i^A, a_j^Bを行ったとき、その選択が一致する。すなわち、

$$a_i^A = a_j^B \tag{3.2}$$

である場合を合意の形成と定義する。

エージェントが環境から完全な情報を得られる場合、各エージェントは全く同じ入力を受けるため、これに対し全く同じ決定的(非確率的)アルゴリズムを用いて選択を行えば合意が成立することは明らかである。しかし、外乱などの影響によって完全な観測が期待できない場合や、他のエージェントの選択アルゴリズムが未知である場合、その未知の選択アルゴリズムを推定する必要がある。

そのような問題に対しては、一定時間内の試行の間の選択頻度から、選択が一致する期待値を得ることができる、確率的な選択アルゴリズムが有効であると考えられる。つまり、他のエージェントの未知のアルゴリズムに対して、確率的な近似による接近を図るものである。

状態sにおいて確率的な選択を行うために、エージェントkは式(3.1)に示した選択可能集合$\boldsymbol{A}$に対応する確率パラメータの集合

$$P^K(s) = (p_1^k, p_2^k, \cdots, p_n^k) \tag{3.3}$$

に従って選択を行うものとする。

4.強化学習によるフォーカルポイントの獲得

前章における定式化のもとで自律的に合意形成を促進させるためには、問題空間に対して最適な確率パラメータpの獲得、つまり、選択可能集合$\boldsymbol{A}$に対応する強化値の集合$\boldsymbol{W}$を学習することが必要であり、その手法として経験強化型の強化学習法であるProfit Sharing [4]を採用する。強化学習が有する特徴は、領域に対する設計者の知識を必要とせず、学習主体と環境の相互作用に依存した汎用的な学習アルゴリズムである。

4.1 強化学習

強化学習(Reinforcement Learning)は、動物の適応過程を工学的観点からモデル化した枠組である。学習主体であるエージェントは未知なる環境と直接対時する。環境からの感覚入力(sensor)に対して、行動(action)を選択し、実行に移す。一連の行動に対する結果として、環境からの報酬(reward)が与えられる。

ここで、報酬遅れが伴うこと、報酬に至る各行動の適否についての情報は存在しないことに注意する。エージェントと環境の相互作用の繰り返しを通じて、獲得する報酬を最大化するための最適または合理的な政策を学習することが強化学習の目的である。図２に強化学習の枠組を示す。

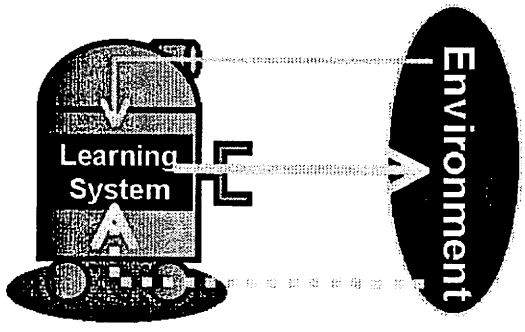


図２：強化学習の枠組

4.2 Profit Sharing

Profit Sharing[4]は、市場経済のアナロジーから生まれたヒューリスティックな方法である。報酬に至る行動系列における感覚入力 x と行動 $a(\in A)$ の対からなるルール系列を記憶しておき、報酬が得られた時点で、系列上のルールを次式に従って強化する学習手法である。

$$\varpi(x_i a_i) \leftarrow \varpi(x_i a_i) + f(\mathrm{r}, i) \qquad (4.3)$$

$\varpi(x_i a_i)$: 行動系列の i 番目の強化値($\in W$)

r: 報酬値，f: 強化関数

宮崎らは、学習過程における行動選択の方法として、ルーレット選択が良い性能を示すことが示された。また、ルーレット選択は、確率的政策を自然に実現する枠組でもあり、局所解から脱出するための手法としても有効である[5]。

Profit Sharing が有する性質として、

- 報酬を経験として直接フィードバックするため、学習に要する時間が短い
- 経験の強化という観点から、不完全な環境情報においてもロバストに機能する

が挙げられる。これらの特徴は、一般的に環境情報の同定が困難であるマルチエージェント環境において、合意形成を促進する手法として有効であると考えられる。学習の方針は、複数のエージェントが存在する環境において、各エージェントの選択が一致した場合に報酬を得られるものとする。このようにして与えられる報酬を手がかりにすることで、フォーカルポイントの自律的な獲得が可能であると考えられる。

5.マルチエージェントと Soccer Server

マルチエージェントシステムの研究モデルとして、サッカーが取り上げられている[2]。サッカーは、競合相手が存在し、動的に変化する環境において、目的を達成するために複数のエージェントが強調するという意味で、マルチエージェントシステムに要求される特徴を含んだ問題となっている。

実際のサッカーゲームでは、味方プレイヤ同士での合意の形成が重要となる。つまり、同じフィールド上に存在する敵プレイヤに如何に察知されることなく、これから展開される作戦を味方に伝達し、実行するかということである。その方法として、事前の合意を形成する、サインを出す、アイコンタクトを用いる、などが挙げられる。これらは、会話などの、明示的な通信が著しく制限された状況における合意の形成手法であり、協調問題の解決には不可欠な要素であろう。

このような、サッカーの持つマルチエージェントシステムの興味深い研究領域を損なうことなく、適度に簡単化した形のフレームワークとして提供されているソフトウェアが、Soccer Server[1]である。

Soccer Server は、計算機上に仮想的なサッカーフィールドを提供し、各プレイヤはクライアントとなってそのフィールド上を「走る」、「蹴る」などの制御命令を用いてゲームを行う。このソフトは、実験環境の自由度の高さと RoboCup[2]によって国際的に認知されたことにより、人工知能の関連技術に対する標準的な評価の枠組として定着する可能性がある。

本研究では、Soccer Server を用いて、提案手法の有効性を検証するための具体的な問題設定としての「パス協調問題」をシミュレートし、マルチエージェントシステムのための、協調アルゴリズムの評価を行う。

5.1 パス協調問題

パス問題を設計するにあたり、実験環境が有する制約条件を以下に列挙する。

- 各エージェントは自身が知覚できる環境情報のみに基づいて行動する。
- 知覚情報は完全ではなく、雑音の影響を受ける場合がある。
- エージェント同士は直接通信することができない。

これらの制約によって、実験環境は実世界に近い多様性を有することになる。図 3 に示されるパス協調問題を以下のように定義する。

敵陣での敵味方 3 人ずつの状況において、攻撃側のプレ

[1] http://ci.etl.go.jp/~noda/soccer/server/
[2] http://www.robocup.org/

イヤ(エージェント)が以下のタスクを達成することを目的とする。

1. Passer は Receiver へパスを成功させる。
2. Receiver は Passer からのパスを受けとり敵ゴールへのシュートを成功させる。

パス協調問題においては、Passer と Receiver の間での合意形成が重要になる。例えば、敵味方の位置関係によって、味方プレイヤへの直接のパスが有効な場合や、敵プレイヤが存在しない場所へ味方を走り込ませるためのスルーパスが有効な場合がある。どちらの場合も Passer と Receiver の間で「誰に(あるいはどこに)パスを出すのか」という合意の形成がなされていなければパスの成功は見込めない。また、パスの成功をシュートの成功に結びつけるためには、できるだけゴールに近く、なおかつ守備プレイヤから離れた場所でパスを受ける必要がある。

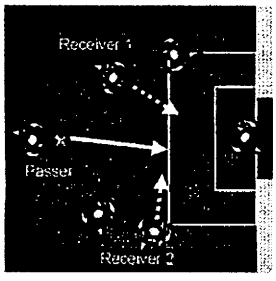


図 3：パス協調問題

このような理由に基づいて、タスク達成のために攻撃側プレイヤが獲得すべきフォーカルポイントは「フィールド上における特定の場所」と考え、サッカーエージェントの設計を行う。

5.2 サッカーエージェントの設計

5.2.1 エージェントの行動空間

状態空間が計算機のメモリに収まることが前提となる強化学習を、座標系などのパラメータの多くが連続値で表現される Soccer Server 環境に適用するために、空間の離散化を行う。

図4に12×16の格子空間に分割されたフィールドを示す。なお、パス協調というサッカーにおける問題の局所性から、フィールドは守備側の陣地に限定する。エージェントが移動する場合や対象に何らかの処理を行う場合、格子空間を単位として行動する(以降、この格子空間をセルと呼ぶ)。ただし、自分が位置するセルにボールが存在し、それを追跡しなければならない場合はその限りではない。

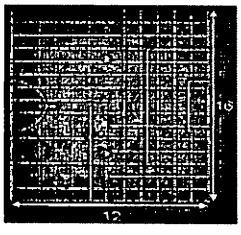


図 4：格子空間に分割されたフィールド

5.2.2 基本行動アルゴリズム

I　攻撃側エージェント(学習主体)

攻撃側エージェントには Passer と Receiver の 2 種類エージェントが存在する。これらのエージェントは、Profit Sharing による強化学習システムが実装されており、内部的な構造には差異はないが、基本的な行動アルゴリズムが異なっている。

エージェントは、選択可能集合 $\boldsymbol{A}$ として図4に示されるような2次元格子空間を持つ。

すなわち、

$$\boldsymbol{A}=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, (\mathrm{m}=16, \mathrm{n}=12) \qquad (5.1)$$

となる。

また、エージェントは図4で示された環境における観測と行動から得られる報酬を通じて、ひとつひとつのセルに対する評価を、エージェントが有する強化値の分布として写像する。つまり、図5に示すように、問題空間の場の状況をエージェントが共同注視する事によって問題解決を図るものと考え、格子空間に分割されたフィールドに対する強化値の分布を、学習によってエージェント間で接近させる事で合意を形成するものである。

図6に示すように、Passer は強化値のルーレット選択によりパスを出す場所を決定する。Passer はその役割の性質上、ボールを追いかけるなどの移動は行わない。

Receiver の移動位置の決定法を図7に示す。自分の位置を含め、隣接するセルに対応する強化値についてルーレット選択を行うことで、次の移動位置を決定する。

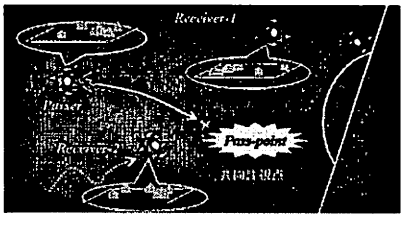


図5：強化値の分布状態の接近による協調行動

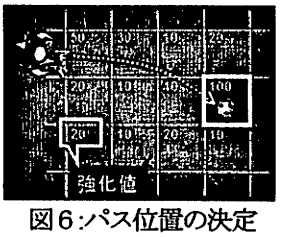


図6：パス位置の決定

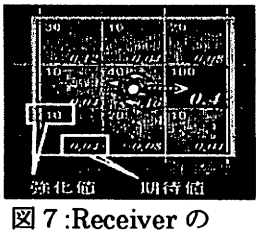


図7：Receiver の移動位置の決定

II　守備側エージェント

守備側のエージェントは学習を行わず、シンプルな守備行動をとる。その行動アルゴリズムを以下に示す。

i 初期位置にしたがって配置につく。
ii 観測物からボールを探し、常にボールの方向を向く。
iii 自分とボールまでの距離がある閾値以下ならば、ボールに接近する。
iv ボールを捕獲したら、相手陣地へ向かって蹴りだす。

5.2.3　行動系列の強化と観察による強化

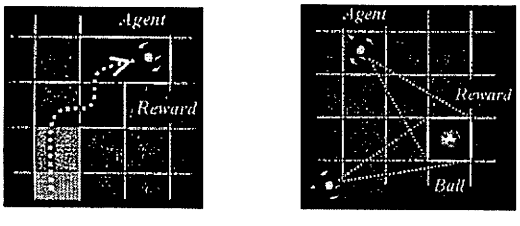
本研究では、学習主体であるエージェントに対して行われる強化の方法として、以下の二種類の方法を用いる。

1.報酬に至る行動系列の強化
2.観察による強化

1.は一般的な強化学習で用いられる方法であり、エージェントが実際に行動した系列だけを強化する(図8(a))。これはエージェントにとって「自信」を動機とする学習である。2.はエージェントの観測情報を積極的に利用することを目的として導入する強化の方法である。

これは、エージェントが受ける観測情報から現在の状態に関する文脈を推論することで、現時点では実行不能であるが、有効であると思われる政策を強化するものである[6]。

サッカーを例に挙げると、Receiver が「Passer の蹴ったボールが、味方がいない場所に転がっている」という観測情報を得た場合、ボールを拾うためには現在の位置よりも、その場所がより適していると推論することができる(図8(b))。このような、有効と思われる行動を、観測情報から推論することによって強化する手法は、人間が行っているところの「後悔」という振る舞いに相当する学習であり、エージェント間で合意形成を図る上でも有効であると考えられる。



(a)行動系列の強化　(b)観察による強化

図8：強化の方法

5.2.4　強化関数

次に、Profit Sharing に適用する強化関数を設計する。格子空間に分割された環境をエージェントの選択可能集合 A として表現したことにより、注視点として選択された要素 a_{ij} は、周辺の要素に対しても影響を及ぼすと考えられる。そのような特徴を持つ強化関数として、以下に示すポテンシャル関数を採用する。この関数は、座標 (x_p, y_p) が最大の強化を受け、その点から遠ざかるほど受ける強化は弱くなるという特徴を持つ。

$$P_{xy}(t+1)=P_{xy}(t)+\alpha\exp\{-\beta\{(x-x_p)^2+(y-y_p)^2\}\} \qquad (5.2)$$

x_p, y_p：報酬を与えられた場所
α：報酬。ただし、行動系列の強化の場合には項比 d の等比減少関数
β：報酬の拡散係数

また、試行毎に全体の強化値を一定の減衰率 γ によって減衰することにより、無効ルールの抑制と、長時間の試行に伴う強化値のオーバーフローを回避する。

6. 計算機実験

6.1　実験1：フォーカルポイント学習実験

パス協調問題の実験手順を以下に示す。

Step 1：攻撃側エージェント(Passer,Receiver1,2)、守備側エージェント(Defender1,2,3)を定められた位置に配置する。
Step 2：Passer はパスを行い、Receiver はボールを拾ったらゴールへシュートする。各エージェントは、観測されたボールや味方エージェントの位置を強化する。
Step 3：プレーが成功した場合、エージェントの行動系列を強化関数にしたがって強化する。
Step 4：試行回数が終了条件を満たさなければ Step 1 へ。

なお、エージェントの位置関係は固定であり、1試行を約10秒間として5000回学習させる。

強化関数のパラメータを表1に、実験で用いたエージェントの初期配置を図9に示す。

表1：強化関数のパラメータ

報酬 α	行動系列の強化	pass/shoot/goal: 50 (初項50 項比0.4 の等比減少関数)
	観測による強化	20
拡散係数 β	0.08	
減衰率 γ	0.99	

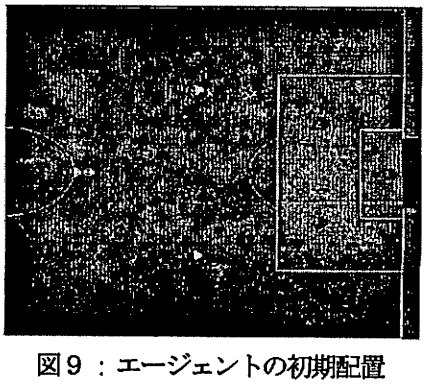


図9：エージェントの初期配置

§ 実験結果および考察

学習後の攻撃側エージェントについて、獲得された強化値の分布図を図１０に示す。さらに、各エージェントについて強化値が 500 以上の値を示す部分を抜きだし、それを重ね合わせたものを図１１に示す。

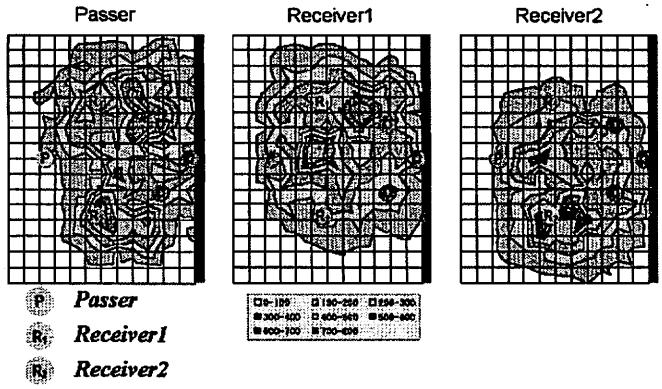


図１０：攻撃側エージェントの強化分布図

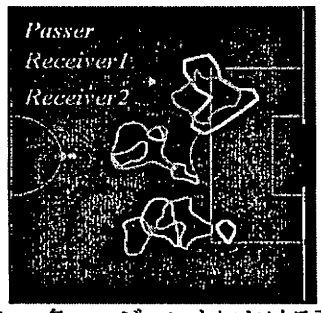


図１１：各エージェントにおける強化値
500 以上の分布の重なり

①Passer

全体として強化値の勾配の差が激しく、分布の様子は複雑である。その一方、500~700 という高い値を持つ強化値は、Passer からゴールの方向に対して左右に分かれており、明確な偏りが観察できる。これらの位置は各 Receiver の初期位置から若干ゴール方向寄りであることから、Receiver からそれほど遠くない位置へパスを行う傾向があると考えられる。それ以外では、両 Passer と、ゴールのほぼ中間点付近の強化値が高い値を示しており、ゴール正面方向へのパスも試みていたことが分かる。

②Receiver1 , 2

Passer の分布と比べて強化値の勾配は緩やかであり、各々２つのピークを持つランドスケープとなっている。例えば、Receiver1 においては、敵エージェントと、初期位置との中間地点に強化値のピークがあり、Passer における強化値のピークとオーバーラップする。Passer 正面付近でのピークは、Passer とゴールの位置関係から、ゴール正面方向へのパスに対して獲得されたと考えられる。Receiver1 はこの 2 つのピークに引き込まれる振る舞いを見せたことが分かる。

Receiver2 の強化値の分布は、Receiver1 の分布と対称な分布が観察できる。

Passer ～ Receiver 間での強化値分布の一部にオーバーラップが見られたことより、パス協調問題に対して合意形成を図る役割を持つことが確認できた。また、どのエージェントもゴール付近での強化値は低い値を示していることから、ゴールに近い位置でのプレーは敵エージェントに妨害されやすいため有効ではないという知識を得たものと考えられる。

6.2 実験２：タスク達成率の比較実験

提案手法の性能評価実験として、パス協調問題のタスク達成率について以下のエージェントとの比較を行う。

① 合意形成を行わないエージェント
Passer は前方のランダムな位置に対してパスを行う。Receiver はそのボールを追跡し、シュートする。

② 固定の知識によって合意形成を行うエージェント
設計者が用意した６通りの固定的なパスコースに対して、合意形成の学習を行う。

§ 実験結果および考察

実験１と同様の手順で5000 回の試行を行った結果を表２に示す。合意形成を行わないエージェント(no-agreement-method)は、ランダムなパスに対して Receiver が追従できないために、達成率の向上は見られなかった。固定の知識によって合意形成を行うエージェント(fixed-method)は、決められたパスコースに対する合意形成の学習は行われたが、Receiver が集中してしまい、デッドロックが生じる場面が見られた。提案手法(proposed-method)は、試行初期では他の 2 つの手法と比較して達成率が低かったが、試行を繰り返すことでパスが成功するようになり、最終的に 29%程度の達成率を示した。

表２：5000 回試行におけるタスク達成率

no-agreement-method	17.36%
fixed-method	20.90%
proposed-method	**28.92%**

7. 結論

本論文では、マルチエージェント協調問題に対するアプローチとして、強化学習を用いた共同注視点に基づく合意形成による問題解決の枠組を提案し、サッカーにおけるパス協調問題に適用することで提案手法の有効性を検証した。

本研究では、強化学習における強化の方法として、エージェントが自らの行動によって行われる強化と、処理の対象物(他のエージェントなど)の観測によって行われる強化の２つの方法を用いた。

観測による強化の概念を、より一般的な社会における人間の行為と照らし合わせると、注意(attention)[国吉 95]に相当すると考えられる。本来、行動主体が観測物から注意する対象を決定する場合、状況の文脈から自律的に判断することが自然であるが、本研究ではパス協調問題というタスクに依存して、注意すべき対象をボールや味方エージェントに限定したため、そのような枠組を得るには至ってい

ない。また、実験によって得られた共同注視点がエージェントの初期配置配置を変化させた場合、有効に機能するかは明らかではない。今後はそのような課題に留意し、動主体が、環境の変化に対してより柔軟に合意形成を図るための枠組について考察する必要があるだろう。

謝辞

本研究の一部は、財団法人テレコム先端技術研究支援センターの支援により実施した。

文献

[1] 本村陽一:確率モデルの学習によるエージェント間の合意形成，MACC '97, (1997).

[2] Noda, I. and Matsubara, H : Soccer Server and Researches on Multi-Agent Systems, IROS-96, pp. 1-7, (1996).

[3] Schelling, T.C. : A Strategy of Conflict, Oxford University Press, (1963).

[4] Grefenstette, J.J. : Credit Assignment in Rule Discovery Systems Based on Genetic Algorithms, *Machine Learning*, Vol. 3, pp. 225-245, (1988).

[5] 荒井幸代, 宮崎和光, 小林重信: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. 4, pp. 609-617, (1998).

[6] 安藤友人:サッカーエージェントにおける強化学習を用いたポジショニング, MACC ' 97, (1997).