RoboCupサッカーエージェントのための経験的知識を利用する機械学習手法

星野 孝総‡　亀井 且有‡
立命館大学理工学部‡

1　はじめに

RobuCupは，2050年にヒューマノイド型ロボットで人間のサッカー世界チャンピオンに公式ルールで勝利することを目標とするプロジェクトである[4]．中でもシミュレーションリーグは，エージェント間の協調行動の獲得や学習など，情報科学の幅広い分野の研究が可能なリーグである[3][7]．シミュレーションリーグ[5]には，プレイヤーのポジショニングを強化学習を用いて学習するチームや，遺伝的アルゴリズムを用いて処理を自動プログラミングするチーム[2]など，さまざまな研究がある．われわれは，特徴的な入力に対しファジィルールを用いて行動決定するプレーヤーを提案した[1]．言語ラベルを持つファジィルールを用いてプレイヤーを設計できるため，プレイヤーの設計が容易におこなえるようになった．そこで，本論文では経験的知識を学習できる機械学習の一手法である強化学習を用い，ファジイルールのチューニングを試みる．また，人がファジィルールを設計して作成したプレイヤーと対戦実験を通して強化学習を用いたプレイヤーの有効性を検証する．

2　RoboCupシミュレーション

RoboCupシミュレーションは，サーバ·クライアント方式を採用している．作業を依頼する方をクライアントといい，作業を処理する方をサーバという．RoboCupシミュレーションでは，ボールやプレイヤーの動きを処理するサッカーサーバが存在する．それに対し，サッカーサーバーから取得した情報をもとにプレイヤーの行動を決定するプレイヤークライアントが存在し，サッカーサーバに行動のコマンドを送信することが出来る．また，プレイヤークライアントがネットワークなどを使って，他のプレイヤークライアントと情報を直接共有することは，原則として禁止されている．

各チーム11プレーヤークライアント，計22プレイヤークライアントがサッカーサーバと通信を行うことにより試合を進められる．試合内容はサッカーモニタと呼ばれるプログラムを通して見ることができる．サッカーモニタを図1に示す．

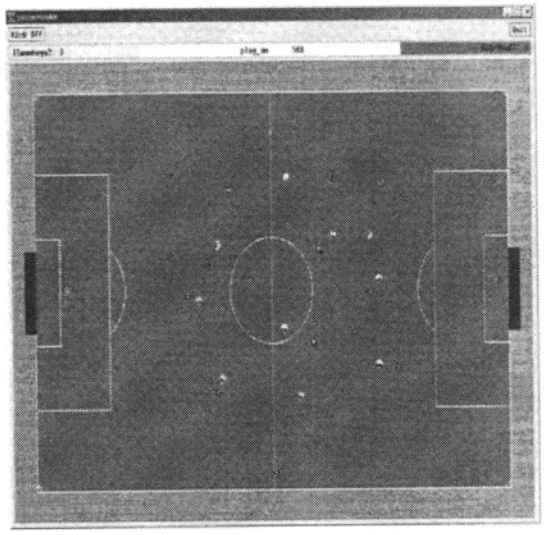


図 1: サッカーモニタ

2.1　サッカーサーバ

サッカーサーバには，ステップ時間が設定されており，離散時間で処理が進められる．ある時間において，各プレイヤーとボールの位置，向き，スピードがわかっている．ここで，各プレイヤークライアントから送られてきた行動のコマンドから，サッカーサーバは次のステップ時間における各プレイヤーとボールの位置等の情報を計算する．これらはステップ時間毎に計算される．サッカーサーバとプレイヤークライアントの入出力関係を図2に示す．その他にサッカーサーバには，フィールドの環境やプレイヤーの能力を設定することが可能であり，審判の機能も備えている．(以下，サッカーサーバをサーバとする．)

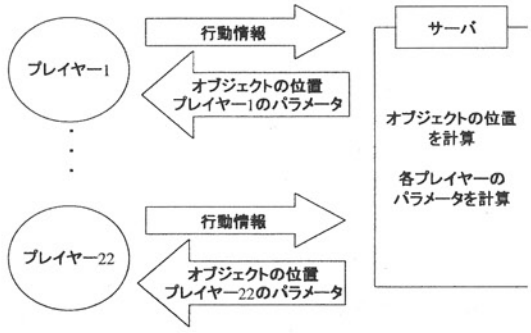


図 2: サッカーサーバの入出力関係

2.2 プレイヤークライアント

プレイヤークライアントがサーバから受け取る情報は，視野情報，聴覚情報，感覚情報に分けられる．

- 視野情報

  プレイヤーは，図3に示すような半径 visible_distance の円状の視野と，角度 view_angle のコーン状の視野を持っている．プレイヤーはこの視野内に存在しているプレイヤーの位置，向き，スピード，チーム，番号，及びボールの位置，向き，スピード等の情報を得ることができる．但し，プレイヤーは，視野内の情報をすべて無条件に得ることはできない．unum_far_length を超えると，プレイヤーの番号が見えない場合が生じ，unum_too_far_length を超えると，プレイヤーの番号が完全に見えなくなる．同様に，team_far_length を超えると，プレイヤーの所属チームが見えない場合が生じ，team_too_far_length を超えると，プレイヤーの所属チームが完全に見えなくなる．これらの条件に従って，サーバから各プレイヤーに各プレイヤーとボールの情報が視野情報として送られてくる．

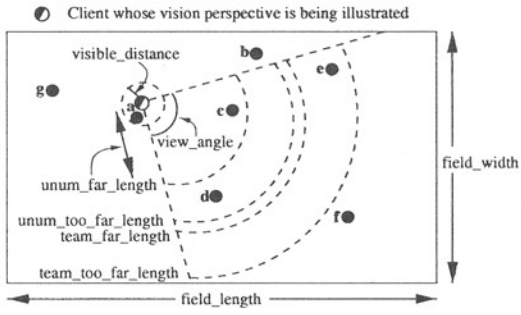


図 3: プレイヤーの視野情報

- 聴覚情報

  視野情報は，審判の声や，他のプレイヤーからの声である．プレイヤー間で直接通信を行うことは禁止されている．このため，プレイヤーは「声を出す」という行動をサーバに送り，他のプレイヤーはそれを聴覚情報としてサーバから受け取ることにより通信を行う．

- 感覚情報

  感覚情報は，スタミナ, 回復力など，各プレイヤーが持つパラメータである．

プレイヤーがサーバへ送る情報は 5 種類の基本行動である．表 1 に基本行動とその意味を示す．

表 1: 行動とその意味

行動	意味
kick	ボールを蹴る
turn	体の向きを変える
dash	走る
turn-neck	視野情報の向きを変える
say	声を出す

プレイヤーの入出力関係を図 4 に示す．

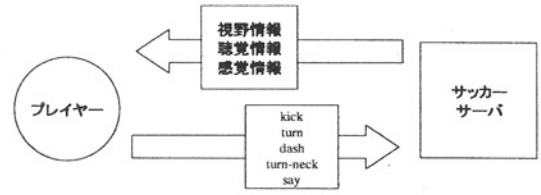


図 4: サッカープレイヤーの入出力関係

本研究では，基本行動を組み合わせた行動を連続行動として定義している．例えば「ドリブル」は，基本行動である「kick」と「dash」を組み合わせた連続行動である．連続行動には，一定ステップ連続して実行される行動がある．Carnegie Mellon University ではパス，ドリブル等の個人技ルーチンをライブラリとして公開している．ソースコードは C++が使われている．このライブラリは CMUnited99(以下，CMUnited99 を CMU とする) と呼ばれており，多くのチームのライブラリとして用いられている．本研究では，パス，ドリブル等の連続行動を CMU を用いてプログラミングを行った．また，CMU にはテスト用プレイヤーとして，ボールを見つけ，ボールに近づき，ボールをゴール方向へ蹴るとい

う単純な動作を行うサンプルプレイヤーが用意されている．

2.3 コーチクライアント

コーチクライアントは，プレイヤーを開発中に，試合を制御するために作られた機能である．コーチクライアントは次のような機能を持っている．

- free_kick，kick_off 等のプレイモードを変更できる．
- 聴覚情報を広めることができる．
- プレイヤーとボールをフィールド上のすべての位置へ移動できる．
- フィールド上の全てのプレイヤーとボールの位置情報を受け取ることができる．

本研究において，ボールがある一定の時間動かなくなった場合，ボールとプレイヤーの位置をゲーム開始時の状態に移動させるために，コーチクライアントを用いている．本研究の実験では野田五十樹氏によって作られた，コーチクライアント libsclient を用いている [9, 10]. libsclient はサーバから情報を受け取る部分と，プレイヤーの基本行動をサーバに送る部分だけが，ライブラリ化されているコーチプログラムである．ソースコードは C 言語が使われている．

3 ファジィ推論

ファジィ理論は，1965 年に Zadeh によりファジィ集合という形で提唱されてた．言語的に内在するあいまい性を取り扱った理論としてまた，Zadeh は続いてファジィアルゴリズム，ファジィ推論の論文を発表し，ファジィ理論の基礎的な理論ができあがった．ファジィ推論は入力に対して，複数のファジィルールが同時に発火することにより，推論結果を得る [6]．ここでファジィルールとは，人間の知識を if-then の形で表現したルールである．ファジィルールは，言語値に対応したファジィラベルで表現されているため，システムを言語的に分かりやすく記述できるというメリットがある．その反面，ファジィ推論システムの設計において，推論ルールの獲得やメンバーシップ関数のチューニングが困難であるという問題がある．

ファジィ推論の推論法はいくつか存在するが，ここでは簡略推論法について説明する．簡略推論法のファジィルールは，前件部にファジィ集合を用いている．そして，後件部に実数値を用いている．このため，ファジィルールは式 (1) のように表される．

$$\begin{cases} R^1: \text{if } x_1 \text{ is } A_{11} \text{ and } x_2 \text{ is } A_{12} \text{ then } y = b_1 \\ R^2: \text{if } x_1 \text{ is } A_{21} \text{ and } x_2 \text{ is } A_{22} \text{ then } y = b_2 \\ \vdots \\ R^i: \text{if } x_1 \text{ is } A_{i1} \text{ and } x_2 \text{ is } A_{i2} \text{ then } y = b_i \\ \vdots \\ R^n: \text{if } x_1 \text{ is } A_{n1} \text{ and } x_2 \text{ is } A_{n2} \text{ then } y = b_n \end{cases} \tag{1}$$

ここで x_1，x_2 は，ファジィシステムの入力変数であり，状態量を表す．y はファジィシステムの出力変数である．A_{i1}，A_{i2} は，ファジィ変数であり，b_i は実数値をとる．いま，入力 x_1^*，x_2^* が与えられた場合の計算手順を示す．各ルールの適合度 w_i を min 演算から求める．その式を式 (2) に示す．

$$w_i = \mu A_{i1}(x_1^*) \wedge \mu A_{i2}(x_2^*) \tag{2}$$

ここで，$\mu A_{i1}(x_1)$ と $\mu A_{i2}(x_2)$ は，それぞれ A_{i1}，A_{i2} への帰属度を表すメンバーシップ関数である．帰属度とは，入力値がファジィ変数に属する度合いである．推論結果 y^* を w_i と b_i の重み付き平均から求める．その式を式 (3) に示す．

$$y^* = \frac{\sum_{i=1}^{n} w_i b_i}{\sum_{i=1}^{n} w_i} \tag{3}$$

簡略推論法は他の推論法と比べ，重心計算がないため，計算時間が早いという利点がある．

4 ファジィ推論を用いて行動を決定するプレイヤー (FRP)

4.1 プレイヤーの構成

構成するプレイヤーは，最も近い敵の位置，最も近い味方の位置，ボールの位置，自分の位置，自分の向きを入力とし，連続行動を出力する．出力である連続行動は，ボールを蹴ることが出来る状態で実行できる行動と，そうでない行動に分けることができる．このため，2 つのファジィ制御器を作成する必要が生じる．また，

ボールが視野内に無い場合は，ボールを捜すため，90度左にturnを行う．プレイヤーの入出力と処理手順を図5に示す．

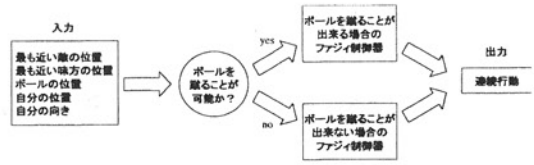


図 5: プレイヤーの処理手順

4.2 ファジィルールの設計

ファジィルールの前件部は，表2にしめすように5つの入力変数を定義する．

表 2: 前件部の入力変数

変数名	情報
Enemy	敵の位置
Teammate	味方の位置
Ball	ボールの位置
MyPosition	自分の位置
MyDir	自分の向き

Enemy, *Teammate*, *Ball* に関するファジィ変数は，図6(a)に示すように，視野を4分割し，Left Far(*LF*), Left Near (*LN*), Right Far (*RF*), Right Near (*RN*) とする．*MyPosition* に関するファジィ変数は，図6(b)に示すようにOpponent Territory (*OT*), Friendly Territory (*FT*) とする．*MyDir* に関するファジィ変数は，Opponent Goal (*OG*), Friendly Goal (*FG*) とする．それぞれのファジィ変数のメンバーシップ関数を図7にしめす．

ファジィルールの後件部の出力変数は各連続行動である．各連続行動に関する値は *Bad*, *Soso*, *Good* で表され，それぞれ 0.0, 0.5, 1.0 の値をとる．出力する連続行動は，表3表4に示すように6種類ある．

これらの前件部の組み合わせは敵，見方，ボールのそれぞれが視野内に無い場合も含めると，500とおり存在し，計500個のファジィルールでシステムが構成されている．例としてボールを蹴ることが出来ない状態における91番目のファジィルール R^{91} を式(4)に示す．

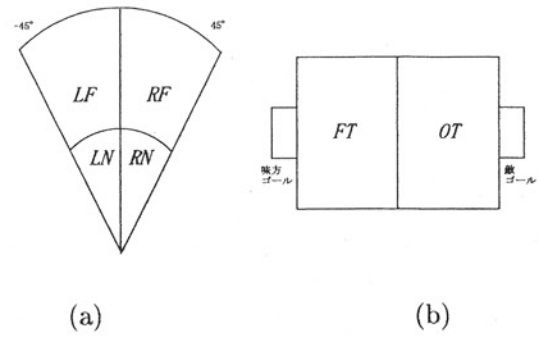


図 6: 視野と位置のファジィ変数

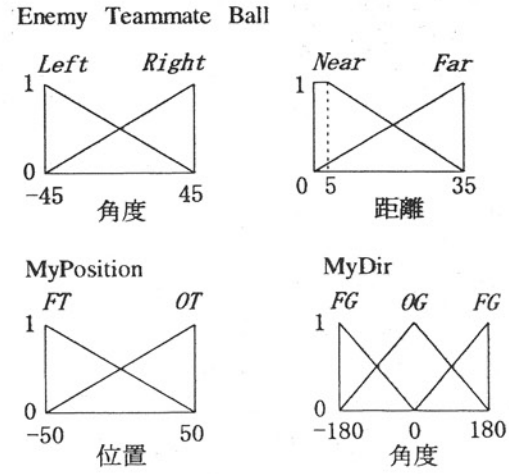


図 7: プレイヤーのメンバーシップ

表 3: ボールが蹴ることが出来る場合の連続行動

変数名	行動
DribbleToGoal	敵ゴール方向へドリブル
Pass	見方にパスする
Shot	シュートする

表 4: ボールが蹴ることが出来ない場合の連続行動

変数名	行動
ToGoal	敵ゴール前へ移動
GetBall	ボールを取る
ToSpace	スペースへ移動

R^{91} : if *Enemy* is *LN* and *Teammate* is *RF* and *Ball* is *LN* and *MyPosition* is *FT* and *MyDir* is *OG*
then *GetBall* is *Good* and *ToGoal* is *Bad* and *ToSpace* is *Bad* (4)

4.3 推論法

本研究で用いる推論法は簡略推論法を基にしている．まず，各ルールの前件部の適合度 w_i を min 演算で求める．そして，適合度と後件部から，連続行動に対する重み付き平均 y を計算する．式 (5) に *Pass* を例として計算方法を具体的に示す．ここで，$Pass_i$ はルール R^i の連続行動 *Pass* における値 (*Bad*,*Soso*,*Good*) を表している．

$$y_{Pass} = \frac{\sum_{i=1} w_i Pass_i}{\sum_{i=1} w_i} \tag{5}$$

各連続行動ごとに重み付き平均を求め，重み付き平均が最も高い連続行動が選択される．

5 強化学習

強化学習は，報酬・罰を手がかりとして，環境に適した行動を強化する学習である．ここで，ある環境の中で行動を起こすエージェントを想定する．エージェントは単位ステップにおいて，環境から状態を感覚入力として受け取り，行動を決定する．決定された行動に対して，エージェントは環境から報酬あるいは罰を与えられる．報酬と罰を強化信号とする．強化学習は，状態認識器，行動選択器，学習器の 3 つのユニットから構成されている．状態認識器は状態を認識して，政策候補の集合を生成し，行動選択器に送る．状態認識器から送られた政策候補の集合から評価値の大きい行動を選択して環境に出力する．この政策により，状態が遷移し，遷移先状態が報酬・罰の条件を満たしているとき，環境は報酬・罰を学習器に与える．学習器は，報酬・罰に従って政策に関する評価値を更新する．

学習の目的は，報酬の総和を未来にわたり最大化することである．現在から未来にわたる報酬の総和を遷移先状態の評価値とし，式 (6) で与える．

$$V = \sum_{i=t}^{\infty} \gamma^{i-t} \cdot r_i \tag{6}$$

ここで，V は状態の評価値であり，r_i は時刻 i で与えられた報酬である．また，γ は割引率と呼ばれる定数で，$0 \leq \gamma \leq 1$ の範囲をとる．しかし，未来の報酬は観測不可能である．そこで，学習器では，式 (7) に示すように，離散時間における評価値を更新する．

$$V_t \leftarrow V_t + f(r) \tag{7}$$

ここで V_t は時間 t における評価値である．f は強化関数と呼ばれ，時刻 t における報酬 r_t の関数である．また，強化学習では新たな政策を開始してから，報酬を受け取るまでをエピソードと言い，離散時間をステップとする．強化学習の実現方法として，ProfitSaring 法，Temporal Difference Method(TD 法)，Q-Learning 等のアルゴリズムが提案されている．本研究では経験強化型強化学習の代表的手法である ProfitSharing 法を扱う．Profitsharing 法の強化関数を式 (8) に示す．

$$f(r) = r\gamma^{T-t} \qquad 0 < \gamma < 1 \tag{8}$$

ここで，T は報酬を得た時刻，t は政策を決定した時刻である．ProfitSharing 法は式 (8) に従い，過去に経験した全ての政策の評価値を更新する．

6 強化学習とファジィ推論を用いるプレイヤー (RLP)

RLP は最も近い敵の位置，最も近い味方の位置，ボールの位置，自分の位置，自分の向きを入力とし，ファジィ推論を行うことにより連続行動を出力する．そして，ProfitSharing 法を用いて，ファジィルールの後件部の値を更新することにより，学習を行う．つまり，連続行動が政策にあてはまり，後件部の値がその評価値にあてはまる．以下に具体的な学習手順を説明する．

kickoff を行ってから，得点又は失点するまでの期間を，ひとつのエピソードとする．連続行動には一定時間連続して実行される行動があるため，RoboCup の単位ステップをそのままステップとして用いると，同じ連続行動が何度も強化されるという不具合が生じる．そこで，情報を得たときに前回と異なった連続行動を出力するまでの連続行動を強化学習の対象とする政策とみなして学習する．エピソードは，敵味方どちらかに得点

したときとし，自チームが得点した場合には報酬が与えられ，失点した場合には罰が与えられる．選択された連続行動 (政策) に対する強化関数 $f(r)$ を，式 (8) の ProfitSharing 法を用いて算出する．連続行動を選択する際に発火したファジィルールの後件部を，$f(r)$ と入力に対するファジィルールの帰属度 w を用いて更新する．ここで，ある時刻で「見方にパスする」が選択されたときのルール i における $Pass_i$ の更新式を式 (9) に与える．

$$Pass_i \leftarrow \alpha Pass_i + (1-\alpha) w f(r) \qquad (9)$$

ここで α は学習率である．また，後件部の初期値には，ファジィ推論を行うプレイヤーのファジィルールを用いている．RLP の概念図を図 8 に示す．

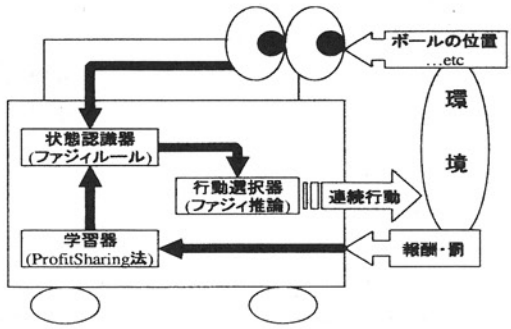


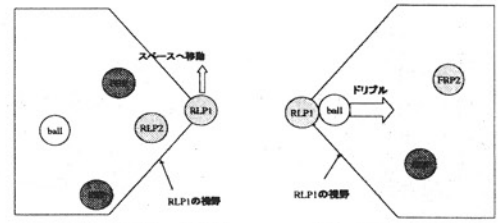
図 8: RLP の概念図

7 対戦シミュレーション実験

7.1 RLP VS FRP(2 対 2)

強化学習を用いたプレイヤー RLP と，筆者らがあらかじめ設計した FRP との対戦シミュレーション実験を行った．対戦人数は 2 対 2 とし，RLP のプレイヤーをそれぞれ RLP1，RLP2 とする．学習のパラメータは，$\alpha = 0.9$，$\gamma = 0.95$ の値をとり，報酬を得た場合には $r = 1$，罰を得た場合には $r = -1$ とする．得失点が 8000 点になったときに試合を終了した．得失点が 2000 点になったときと，得失点が 6000 点になったときにおける，RLP の動作及びファジィルールを検証する．

- 得失点 2000 点のとき
  RLP1 がボールを蹴ることができず，視野内に FRP が 2 体存在する場合では，図 9(a) に示すように，ボールを取りにいかず「スペースへ移動する」場合が多かった．このため RLP2 のみがボールを取りに行くことになる．その結果，2 対 1 でボール取り合うことになり，ボールを奪われ失点する場面が多く観測された．また，ボールを蹴ることが出来ない場合では，図 9(b) に示すように,「敵ゴール方向へドリブル」を選択していた．プレイヤー単独でボールを保持しているため，競り合いでボールを奪われる場面が多く見られた．



(a) プレイヤー単独　(b) ボール保持

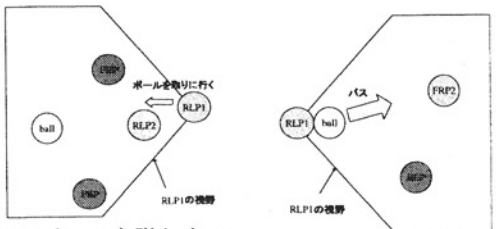
図 9: 2000 点時のプレイヤーの動き

- 得失点 6000 点のとき
  2000 点のときと同様に RLP1 がボールを蹴ることができない場合，図 10(a) に示すように，視野内の FRP と RLP2 に対抗して RLP 2 体でボールを取りに行くようになり，その結果 2 対 2 で競り合う場面が観測され，互角に戦っている．また，RLP1 がボールを蹴れる場合では，図 10(b) に示すように，視野内の RLP2 にパスする場面が多く観測された．RLP2 のファジィルールを調べると，この条件では「パス」の値が高いファジィルールが多く学習されていた．また RLP2 が近くにボールを発見した場合は，ボールを取りに行き，ボールを遠くに発見した場合は，敵ゴールへ向かう行動を選択している．上記の RLP1 と RLP2 の観測された動きにから，RLP1 がミッドフィルダー，RLP2 がフォワードのような役割をするルールが学習されたことがわかる．

この実験では，2 体のみで学習したため，2 種類の役割に学習されたとも考えられる．そこで，多種類の役割を学習させるためにプレイヤーを 4 体に増やし，4 対 4 の対戦で実験を行った．

7.2 RLP VS FRP(4 対 4)

対戦は 4 対 4 とし，RLP のプレイヤーをそれぞれ RLP1，RLP2，RLP3，RLP4 とする．4 対 4 の実験で



(a) ボールを蹴れない場合　(b) ボールを蹴れる場合

図 10: 6000 点時のプレイヤーの動き

は，多くの敵に囲まれる場合が想定される．この場合，計算時間がかかりコマンドをロスする恐れがある．また多くの場合，スペースは見方ゴール方向になることが判っている．そこで，計算時間のかかる「スペースへ移動」を「味方ゴール前へ移動」に変更した．変更した連続行動を表 5 に示す．

表 5: 変更した連続行動

変数	行動
ToGoal	敵ゴール前へ移動
GetBall	ボールを取る
ToFriendlyGoal	見方ゴール前へ移動

他の設定，及び学習のパラメータは先の実験と同様に設定した．得失点が 8000 点になったときに試合を終了した．得失点が 3500 点になったときと，得失点が 7000 点になったときの，RLP の動作，及びファジィルールを検証する．

- 得失点 3500 点のとき
  図 11 に示すように，RLP1 と RLP3 が敵ゴール前へ移動し，RLP2 と RLP4 がボールを取りに行く場面が多く観測された．このため，2 対 4 の競り合いとなり，ボールを奪われ失点する結果となっている．

- 得失点 7000 点のとき
  図 12 に示すように，RLP1 が敵ゴール前へ移動し，RLP2，RLP3，RLP4 がボールを追う場面が多く観測された．RLP2，RLP3，RLP4 はボールを奪い，RLP1 へボールをパスする．そして，RLP1 はボールを受け取り，シュートを行うことで得点する．この得点パターンが多く観測された．また，RLP4 がボールに向かうときに，RLP1 とボールの間に入る場合がある，この場合，先にボールを

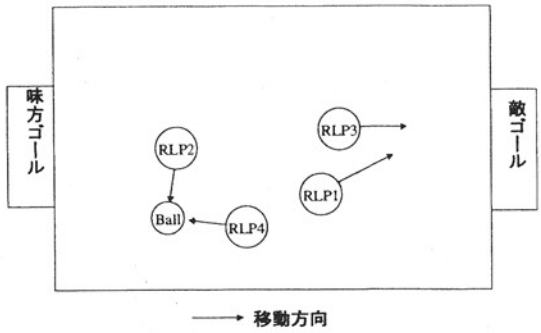


図 11: RLP の動作 (3500 点時)

奪った RLP2 もしくは RLP3 が RLP4 にパスし，さらにゴール前へ移動している RLP1 へパスをするという場面が多く観測された．パスの中継を行った RLP4 のファジィルールは，味方の位置が近い場合に「敵ゴール前へ移動」の値が高く，味方の位置が遠い場合に「ボールを取りに行く」の値が高くなるように学習されていた．これにより，プレイヤーが取得した情報に応じて選択する行動を変えるプレイヤーが強化学習によって習得できることがわかった．

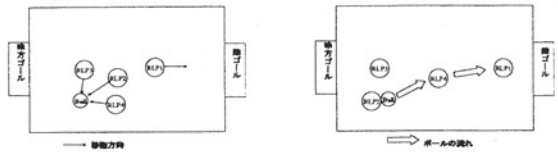


図 12: RLP の動作とパスの経路 (7000 点時)

得失点が 7000 点になったとき，ゴール前へ移動するプレイヤーが 1 体，ボールを取りに行くプレイヤーが 3 人いることが確認できた．また，ボールを取りに行くプレイヤーのうち 1 体が状態に応じてパスの中継を行う場面も見られた．ファジィルールの後件部の値を見ると，2 対 2 の場合は味方の動きに応じた連続行動が選択されるようなファジィルールは学習されなかったが，4 対 4 の場合ではそのような学習できた．以上のシミュレーション実験により，RLP は学習を行うことで，FLP よりも強くなることがわかった．

8 おわりに

本論文では，特徴的な入力が得られた際にとる行動をファジィルールで記述し，ファジィ推論により行動を決定するプレイヤーを示した．次に，強化学習を用いて

ファジイルールの後件部をチューニングするプレイヤーを提案し，それらを対戦させ，その得失点から有効性を示した．学習することで，パスの中継を行うプレイヤー，フォワードの役割をするプレイヤーが観測された．フォワードの役割をするプレイヤーは得られたが，ディフェンダーの役割をするプレイヤーは得られなかった．

今後の課題として，ディフェンダーの役割をするプレイヤーを学習するためのファジィルールの改善と強化学習手法の改善があげられる．また，報酬を与える方法にも工夫する必要があると考えられる．これらへのアプローチとともに，ファジィ変数を進化的に獲得する手法や，ファジィ変数の獲得を自動化する手法にも取り組みたい．

参考文献

[1] 倉多，星野，亀井：強化学習による RoboCup サッカーエージェントの行動獲得に関する研究，第 19 回ファジィシステムシンポジウム，pp.297–300(2003)

[2] 高橋，伊藤：RoboCup ではじめるエージェントプログラミング，共立出版，2001 年

[3] 前田：マルチエージェントロボットにおける協調行動学習のための進化シミュレーション，日本ファジィ学会誌,Vol.13,No.3,pp.281–219(2001)

[4] 浅田：RoboCupSoccer ロボットの行動学習・発達・進化，共立出版，2002 年

[5] 西野：ロボカップ：シミュレーションリーグ，日本ファジィ学会誌,Vol.14,No.6,pp.575–583(2002)

[6] 菅野 道夫：ファジィ制御，日刊工業新聞社，1993 年

[7] 有働，中島，石渕：ファジィ Q 学習によるサッカーエージェントの構築，第 18 回ファジィシステムシンポジウム，pp.89–90(2002)

[8] 河原林，久保，森下，高橋，小高，小倉：ファジィ意思決定によるサッカーパスポイントの決定，第 17 回ファジィシステムシンポジウム，pp.431–434(2001)

[9] http://www.ida.liu.se/ frehe/RoboCup/Libs/, Welcome to the RoboCup simulation library archive

[10] 野田五十樹 , "お気軽サッカーを目指して", 情報処理学会誌 Vol.44 No.09, pp.927–930 (2003)