F-042

サッカーエージェントにおけるスルーパスの強化学習

Reinforcement Learning of through passing in Soccer Agents

田川 諒† 五十嵐 治一†
Ryo Tagawa Harukazu Igarashi

1. はじめに

RoboCup サッカーシミュレーションリーグ 2D [1][2]は，マルチエージェントシステムにおけるエージェント制御のための標準問題として人工知能の研究対象となってきた．本研究では複数エージェント間の協調行動をオンライン的な強化学習により学習させるシステムを開発した．実際に観戦者が試合を観戦しながら報酬を与え，リアルタイムに学習を行わせたところ，10 試合程度で学習が完了し，効果的なスルーパスを多数回出せるようになった．

本研究ではサッカーエージェントのプログラムとしてオープンソースの agent2d を使用したが，実時間での強化学習を可能とするために，ボールを保持したエージェントの探索木の深さを 2 以下に限定し，エピソードも学習エージェントの 2 つの連続行動だけを含むように短く定義した．また，観戦者が自由に報酬を与えられるように，強化学習の手法としてマルコフ性を考慮する必要がない方策勾配法を用いた．さらに，スルーパスが出やすいように探索木中の枝刈り条件を緩和するなどの工夫も行った．本稿ではこれらの方法と実験結果について報告する．

2．サッカーエージェントへの強化学習の適用

2.1 サッカーにおけるゲームアルゴリズムの階層性

一般に，サッカーのようなマルチプレイヤによる団体競技では，(1)個人技(individual skill)，(2)戦術(tactic)，(3)戦略(strategy)，の３つの階層に分けてゲームアルゴリズムを考えることが一般的である．サッカーの例では，(1)にはボールを望みの方向へ高速にキックする技術や，敵プレイヤに奪われないためのドリブル技術，ボールを持ったプレイヤに対する守備プレイヤの１対１の守備技術，ボールのインターセプトの技術などがある．(2)は比較的少人数のプレイヤが協調して行うチームプレイであり，パス回しによるボールのキープ，攻撃のためのリターンパス，スルーパス(through pass)などの技術がある．また，(3)は全体的なプレイヤ配置（フォーメーション），各プレイヤへの役割割当，ゾーンやマンツーマンによる守備方式，戦術の選択・切替などのチームの監督（コーチ）が必要とする技術である．

一方，人工知能の研究においては，マルチエージェントシステムにおけるエージェントの学習方式として強化学習が有力な手段として用いられてきた．RoboCup サッカーシミュレーションリーグ 2D においても，上述の(1)や(2)のレベルを中心に強化学習の適用が試みられてきた．しかし，(1)は個人技なので複数エージェント間の協調行動を取り扱う必要はない．本研究では，複数エージェント間の協調行動の学習に関心があり，上述の(2)のプレイに焦点を絞ることにする．

† 芝浦工業大学工学部情報工学科

(2)のレベルの協調プレイに対する強化学習の適用例としては，Stone らの Keepaway[3](3v2，攻撃側 3 人と守備側 2 人での対戦の意味．以下同様)や Half field offence[4](4v5)の研究，Riedmiller らの複数人の攻撃プレイヤの行動決定の研究[5] (3v4)がある．これらの研究で得られた技術は実際の競技会の中でも用いられた．

2.2 agent2d におけるチェーンアクションの登場

しかし，前章末で述べた研究では，11 人対 11 人のフルゲームではなく，人数を制限した部分ゲームだけを取り扱っていた．かつ，各プレイヤは自分の行動決定だけを与えられた環境の中で条件反射的に学習するだけであり，その場に応じた適切な複数プレイヤが関わる協調行動を計画，実行するという機能はなかった．すなわち，チームメイトは環境の中に含まれており，パスの連携のような協調行動の出現は，各エージェントの単独行動が偶発的に重なった結果に過ぎなかった．それに対し，秋山はフルゲーム中でチームメイトの行動をも含めた行動の連鎖を動的に計画する「チェーンアクション」(chain action)と呼ばれる機能を考案した[6]．この機能は秋山自身がすでに開発していたオープンソースプログラム agent2d (ver. 3.0.0，2010 年公開)へ組み込まれた．agent2d は，日本を中心に現在も多くのチームがベースプログラムとして使用しており，このチェーンアクションの機能を用いている．

2.3 強化学習による局面評価関数の改良

チェーンアクションでは，パス，ドリブル，シュートなどのボール保持者の行動を探索木と局面評価関数を用いて決定する．しかし，agent2d に組み込まれている評価関数はボールとゴール間の距離だけを用いる極めて簡単な関数であった．そこで，谷川らはサッカーの局面評価に関するヒューリスティクスを用いて，複数の評価項の線形和により構成された評価関数を考案し，強化学習による重みの学習を行った[7]．この時，報酬はエピソード中でのボールの移動距離に応じて自動的に与えていたが，3000 試合学習しても学習後のチームが agent2d に勝ち越すことは出来なかった．

学習後のチームが学習前のチームには勝ち越すが，agent2d には勝ち越せない原因の一つは報酬の質にあると考えられた．そこで，田川らはサッカーに関する人間の主観評価を用いて評価関数中の重み係数を決定することを試みた[8]．この研究では， 試合後に局面の静止画を被験者に見せ，局面の評価値と評価関数中の各項の値との相関強度から重みを決定したところ，agent2d に対して最高 59%程度の勝率を持つチームを作ることができた．

2.4 本研究における強化学習システム作成のための基本方針

2.3 の最後で述べた田川らの重み決定方式[8]は，人間の主観評価がサッカーの試合の局面評価に有効であること

を示唆している．しかし，何らかの学習理論に基づいた方法ではなく，重み決定のための処理もバッチ処理的な方法であった．

本研究では強化学習理論に基づいたオンライン的な学習方法を用いる．すなわち，試合中に観戦者が自由に局面の優劣を評価し，その情報を報酬とするオンライン強化学習システムを作成した．強化学習としては方策勾配法[9]を用いた．これにより，報酬のマルコフ性という制限がなくなるので，エピソード中の一連の状態・行動列に対して自由に報酬を与えることができる[10]．また，将来的には，観戦者の試合中の評価や教示を自然言語で行うことを考えている．この際，観戦者の教示内容から正解行動を取りだして，教師有り学習をさせることも計画している．さらに，方策勾配を用いた教師有り学習法[11][12]も提案されており，方策勾配法を用いることには今後の学習内容の範囲を容易に拡張できるという点でメリットがある．

また，人間が報酬を与えるのであれば学習回数を多くとも数十試合程度に減らす必要がある．そこで，5.3 で述べるようにエピソード長を短くし，どの行動や局面に対して報酬が支払われたのか，特定できるようにエピソードを定義した．

3. エージェントの行動決定

3.1 チェーンアクションでの確率的方策の利用

本研究で使用した agent2d は， RoboCup2010 世界大会の優勝チーム「HELIOS」を基に，サンプルプログラムとして公開されているチームプログラムである[13]．基本的な行動戦略は既に実装されており，チーム開発を支援するライブラリも豊富に用意されている．

agent2d の ver3.1.0 以降では，「チェーンアクション」と呼ばれる手法により，ボール保持時における行動の決定を行う[6]．ボール保持者はパスやドリブルといった行動を枝，行動の結果生じる予測状態をノードとする探索木を生成する．各ノードは，評価関数によって優劣が点数化される． agent2d では，点数が最も高いノード(必ずしも葉ノードとは限らない)を最良優先探索によって探索し，そのノードへ至る行動を決定論的に選択する(max 戦略)．ただし，本研究において評価関数の学習を行う際には，次のボルツマン分布で定義される確率的方策を使用した．

$$\pi\left(a\middle|s;\omega\right)\equiv\frac{e^{E_a(a,s;\omega)/T}}{\sum_{x\in A(s)}e^{E_a(x,s;\omega)/T}} \quad (1)$$

ここで，$E_a\,(a,s;\omega)$は局面 s における行動 a の評価値，$A(s)$は局面 s においてエージェントが選択できる行動集合である． 先行研究[7]では，この $E_a\,(a,s;\omega)$の値を，行動 a の実行で得られる局面ノード以下の探索木中で，局面評価値が最も高いノード s_a の局面評価値 $E_s(s_a;\omega)$で置き換えた．すなわち，

$$E_a\left(a,s;\omega\right)=E_s\left(s_a;\omega\right) \quad (2)$$

とした．

図 1 は現在の局面 s からの行動を枝とする探索木の例を示しており，数字はその局面における局面評価関数 $E_s\,(s;\omega)$の値を表している．この例では，行動 a, b, c の行動評価関数 $E_a\,(a,s;\omega)$の値はそれぞれ 80，30，100 となるが，これらの値は局面 s_a，s_b，s_c を局面評価関数 $E_s\,(s;\omega)$により評価した値である．サッカーの場合，オンラインで計画立案と行動決定を行う必要があり，探索木はなるべく小さい方が望ましい．本研究では，スルーパスの学習を目的としているので，パッサーの行うスルーパスの行動とそれを受け取ったレシーバの行動（シュートやドリブルなど）の２つの連続行動だけを計画できれば十分であると考えて，5.の実験では探索木の深さは最大２と設定した．

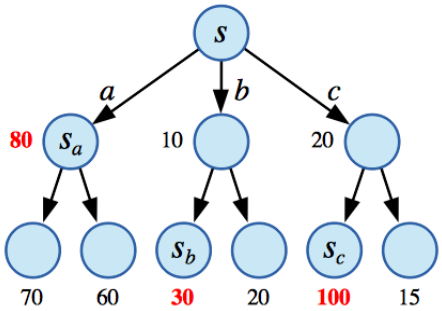

図 1　探索木と局面評価関数の値の例

3.2 局面評価関数

agent2d における局面評価関数は，ボールと敵ゴールの距離のみに依存した単純な関数であった．そこで，谷川らの研究[7]では次の(3)のような評価関数が提案された．

$$E_s\left(s;\omega\right)=U_0\left(s\right)\left[\omega_1U_1\left(s\right)+\omega_2U_2\left(s\right)\right]+\sum_{i=3}^{5}\omega_iU_i\left(s\right) \quad (3)$$

ここで，$U_i\,(s)$ (i=0,1,…,5)は，局面 s を評価する際に有用だと思われる 6 つのヒューリスティクス(先見的知識)を表した関数であり，ω_i (i=1,2,…,5)は重み係数である．(3)の評価関数 $E_s(s;\omega)$は，正/負で絶対値が大きいほど自身の所属チームが優勢/劣勢な局面を表すよう定義している．

(3)の各項の内容を簡単に述べると，$U_0(\in\{-1,0,1\})$は，ボール保持者の所属チームを表している．U_1 はボール保持者と敵の距離，U_2 はボール保持者から見て敵ゴール側にいる敵と味方の人数比，U_3 はボールと両ゴールの距離，U_4 はボールと両チームのプレイヤの距離，U_5 はボール周辺のプレイヤの分布をそれぞれ表している．なお，U_1 と U_2 は[0, 10]，U_3 から U_5 は[-10, 10]の区間内の値を取るように正規化されている[7][8].

4. 方策勾配法による局面評価関数の学習

4.1 学習則

エピソードを定義し，エピソード終了後にエピソード中の状態・行動列を評価して報酬を与える．この報酬の期待値を最大にしたい． (3)を目的関数とするボルツマン分布を（確率的）方策として用いた場合，パラメータ ω の学習則は，強化学習の一種である方策勾配法によると次のように表される[9].

$$\Delta\omega = \varepsilon \cdot r \sum_{t=1}^{L} e_{\omega}(t) \tag{4}$$

$$e_{\omega}(t) \equiv \frac{\partial}{\partial \omega} \ln \pi\left(a(t)\middle|s(t);\omega\right) \tag{5}$$

ただし， $s(t)$は時刻 t における局面， $a(t)$は時刻 t に選択された行動， L はエピソード長，ε は学習係数である．(5)に(2)を代入すると，$e_{\omega}(t)$は次のように表される[7].

$$e_{\omega}(t) = \frac{1}{T}\left[\left(1-\pi\left(a(t)\middle|s(t);\omega\right)\right)\frac{\partial}{\partial \omega}E_s\left(s_{a(t)};\omega\right) - \sum_{x \neq a(t)} \pi\left(x\middle|s(t);\omega\right)\frac{\partial}{\partial \omega}\left(E_s\left(s_x;\omega\right)\right)\right] \tag{6}$$

ここで， (6)の右辺の[ ]内の 2 つの項の符号を考える．第 1 項では，$1-\pi(a(t)|s;\omega) \geqq 0$ なのでエピソード中に学習エージェントが実際に選択した行動 $a(t)$の価値 $E_s(s;\omega)$を高める方向に ω は更新される．第 2 項では，$-\pi(x|s;\omega) \leqq 0$ $(x \neq a(t))$なので $a(t)$以外の行動 x の価値を低下させる方向に ω は更新されることがわかる．

さらに，(3)を(6)へ代入すると，各パラメータ ω_i について最終的に次の学習則を得る．

$$\Delta\omega_i = \varepsilon \cdot r \sum_{t=1}^{L} e_{\omega_i}(t) \tag{7}$$

$$e_{\omega_i}(t) \equiv \frac{\partial}{\partial \omega_i} \ln \pi\left(a(t)\middle|s(t);\omega\right) \tag{8}$$

ただし，(8)は，i=1,2 ならば，

$$e_{\omega_i}(t) = \frac{U_0(s(t))}{T}\left[U_i(s(t)) - \sum_{x \in A(s(t))} \pi\left(x\middle|s(t);\omega\right)U_i(s(t))\right] \tag{9}$$

であり，i=3,4,5 ならば，

$$e_{\omega_i}(t) = \frac{1}{T}\left[U_i(s(t)) - \sum_{x \in A(s(t))} \pi\left(x\middle|s(t);\omega\right)U_i(s(t))\right] \tag{10}$$

である．

なお，(5)や(8)は特徴的適正度(characteristic eligibility)と呼ばれ[9]，強化することがどの程度妥当であるかという強さを表している．特徴的適正度はその定義に方策勾配(policy gradient)を含んでいる．方策関数 $\pi(a;\omega)$は[9]では階層型のニューラルネットワークモデルが用いられていたが，本研究では(1)のようなボルツマン分布関数を用いた．また，報酬が学習パラメータを陽に含まない場合には，状態遷移確率や報酬などの環境モデルと，方策関数に関するマルコフ性を仮定することなく，

$$\frac{\partial}{\partial \omega} E[r] = E\left[r \sum_{t=0}^{L} e_{\omega}(t)\right] \tag{11}$$

が成り立つことが証明されている[10].

4.2 学習システムの作成

本研究では，図 2 に示すような人間による報酬を用いた強化学習システムを作成した．試合観戦者はモニタプログラムを通じて試合を観戦し，投票画面の操作を通じて，評価対象チームのプレイの評価を行う．学習プログラムは，試合観戦者による投票と各プレイヤから得られる情報を重みの学習に用いる．重みの更新は試合中の各エピソードの終了時に， (4)〜(6)の学習則に従って行われる．本研究では，コーチエージェントを利用して，学習計算に必要な情報の収集とエピソードの管理を行い，各プレイヤのパラメータ更新も試合中にリアルタイムで実行するシステムを作成した．

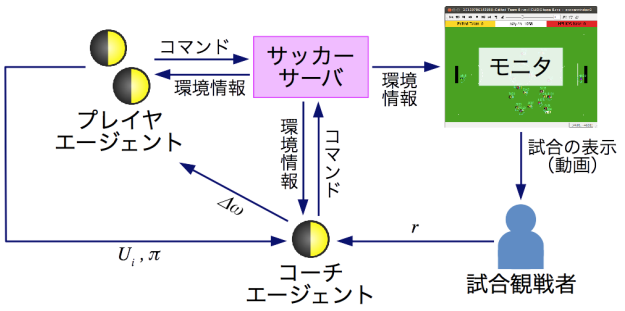

図 2 学習システムの構成

5. 本研究における問題の定式化

5.1 状態 s と行動 a

各プレイヤが得た，敵と味方を合わせた計 22 人のプレイヤとボールの位置情報を状態 s と定義する．また，agent2d に定義されているシュート，パス(ダイレクト，リード， スルー，クロス)，ドリブル(ショート， ロング，シンプル)の 8 種類の行動生成クラスを，次のように探索木の深さに応じて用いた．

- ActGen_StrickCheckPass: ダイレクト・リード・スルーパス(深さ 1)
- ActGen_Cross: クロスパス(深さ 1)
- ActGen_ShortDribble: ショートドリブル(深さ 1)
- ActGen_SelfPass: ロングドリブル(深さ 1)
- ActGen_DirectPass: ダイレクトパス(深さ 2 以上)
- ActGen_SimpleDribble: シンプルドリブル(深さ 2 以上)
- ActGen_Shoot: シュート(深さ 2 以上)

本研究では，これらの 8 種類の行動に加えて，ホールド(現状維持)を加えた全 9 種の行動を行動 a と定義する．

5.2 枝刈りの緩和

agent2d の局面評価関数はボールの x 座標だけを考慮しており，敵にボールを奪われるかどうかという安全性を考慮していない．その代わりに，探索木のノード生成において，局面の安全性を厳しくチェックし，安全性の低いノードを生成しないように枝刈りを行っている．しかし，(3)の局面評価関数中の U_1,U_4,U_5 の項は安全性の評価に関係しており，必ずしもノード生成時に厳格な枝刈り

を行う必要性はない．かえって，チャンスの時にリスクを伴うがチャレンジングな行動が過度に排除されてしまう恐れがある．また，谷川らの研究[7]では，スルーパスの枝刈りを緩和したことが学習における獲得報酬の向上に良い影響を与えたことも指摘されている．

そこで，本研究では，パス(リードパス，スルーパス)とドリブル(ショートドリブル，ロングドリブル)について，agent2d に実装されている味方や敵との位置関係による枝刈りの条件を緩和するよう変更した．具体的な変更点は次の a, b の 2 点である．

a. 敵プレイヤがボールに到達するまでの予測サイクル数を 3 サイクル分だけ長く見積もる (スルーパス，ショートドリブル，ロングドリブルの場合).

b. パスにおけるレシーブ候補地点の生成において，複数のレシーバの候補点が重なった場合の処理を変更する．

上記の変更点 a は，敵が味方より短いサイクル数でボールの位置にたどり着いてしまうような局面を対象とする枝刈りについて条件を緩和するものである．例えば，ボールに最も近い味方がボールを取得するまでに t_m サイクル，敵がボールを取得するまでに t_o サイクル必要だとした場合，agent2d では $t_o > t_m$ が成り立たない局面（ノード）に対して枝刈りを行っていたが，この条件を $t_o > t_m - 3$ に変更することで枝刈りの発生条件を緩和した．

変更点 b については図 3 に示した例を用いて説明する．

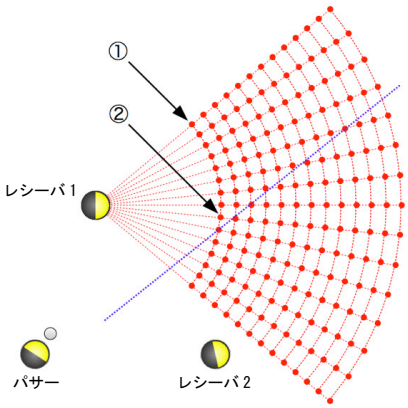

図 3 味方との位置関係によるスルーパスの枝刈りの緩和

図 3 はパサーからレシーバ 1 へのスルーパスの目標地点の生成を示している．一つの目標地点が一つの予測局面（ノード）の生成に対応している．このような場合，agent2d では，通常，図 3 の①で示した目標地点から，等距離にある円周上において時計回りに目標地点の生成を行い，一通り生成が完了した後は 1 段階遠くの円周上にスルーパスの目標地点の生成を続けていく．

ところが，図 3 中央の破線より右下に生成された領域では，目標地点までの距離は，レシーバ 1 よりレシーバ 2 の方が近くなる．このような目標地点についてパサーは，レシーバ 1 にパスを出すよりもレシーバ 2 にパスを出す方が良いと判断するため，レシーバ 1 へのパスとしての生成は行わない．しかも，agent2d の枝刈りでは，そのような目標地点の生成を行うと，以降はそのレシーバに対するパスの生成をすべて打ち切ってしまう．つまり，図 3 の例では，①から②の地点まで生成を行うとレシーバ 1 へのスルーパスの生成を終了してしまい，遠くへのスルーパスが生成されなくなってしまう．そこで，本研究では，このような打ち切りをすることなく，破線より左上側の目標地点が全て生成されるよう枝刈りの緩和を行った．リードパスについてもほぼ同様である．図 4 はこれらの枝刈り処理の変更による実際のノード生成の変化の例を示した画像である．

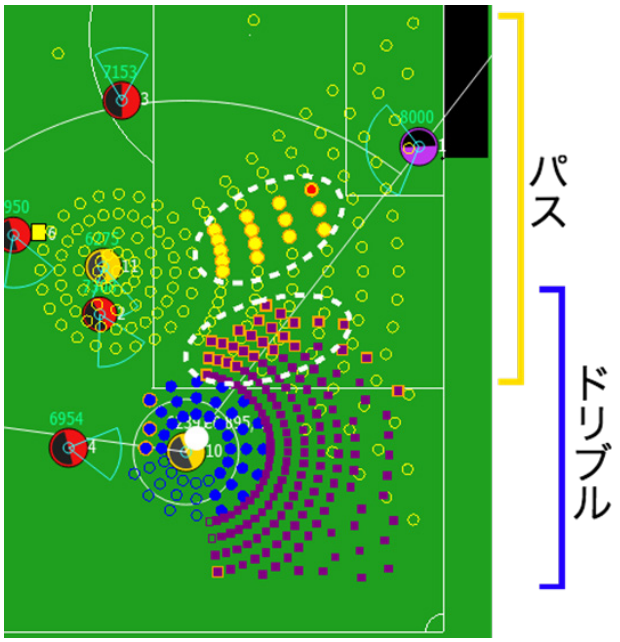

図 4 枝刈りの緩和によるノードの増加例

図 4 では 10 番のプレイヤがボールを持っており，自分自身のドリブルの目標地点と，左上にいる 11 番（円上に配置された丸い記号列の中心に位置）へのパスの目標地点が示されている．内部が塗られていない丸や四角の記号は枝刈りで除去されたノードである．残ったノードが探索木へ加えられる．これらのノード（目標地点）の内，主として太い破線で囲まれた領域のノードが今回の枝刈り条件の緩和により新たに追加されたノードである．

太い破線で囲まれた 2 つの領域の内，上の方の領域はパス(スルーパス)の目標地点を，下の方の領域はドリブル(ロングドリブル)の目標地点を表している．なお，図 4 の局面では，チェーンアクションによる行動選択の結果，最終的には上側の破線領域内の黒丸のノードを選択し，その後得点を決めている．このノードは元の agent2d の枝刈り処理では除去されていたノードである．したがって，図 4 に示した事例は，今回の枝刈り処理の緩和によって挑戦的なスルーパスが生まれた一例と言える．

5.3 エピソードと報酬 r

谷川らの研究[7]においては，自チームのプレイヤがボールを保持してから，相手チームにボールを奪われるか，あるいはボールが場外へ出た場合などプレイが止められるまでの間を「エピソード」として定義してきた．しかし，今回の学習実験では，「試合観戦者が投票した時点

から遡って直前 2 回のボール保持時の学習エージェントの行動とその行動時の局面の組」と定義した．ただし，ここでの行動とは 5.1 で定義したパスやドリブルである．

この理由は次の通りである．今回は人間の観戦者が報酬を与えるので，なるべく学習回数を少なく抑えたい．しかし，エピソードが長いとどの行動決定に報酬を割り当てるのが妥当であるかという問題（報酬割当問題）が生じ，この問題の解決のためにはかなり多数の学習回数を要する．したがって，なるべくエピソード長を短くすることが望ましい．3.1 でも述べたように，今回はチェーンアクションの探索の深さを 2 段として，ボール保持者が連続した 2 回の行動からなる協調行動を計画できるようにした．その計画を評価するには，少なくとも 2 回の連続した行動の結果を観測する必要があり，エピソード長を 2 と定義した．

報酬 r は，エピソード中における試合観戦者の投票によって決定する．投票は「Good」/「Bad」の 2 種類を用意し，それぞれ 1 回投票するごとに報酬が 10 だけ加算/減算される．試合観戦者には，評価対象のチームが得点に繋がる行動を取ったと感じたら「Good」を，得点に繋がらない無駄な行動を取ったと感じたら「Bad」を投票するよう教示した．なお，エピソード終了後 10 サイクル以内ならば何回でも投票ができるようにした．

図 3 にエピソードと報酬の与え方を説明した図を示す．図 3 の例では，観戦者が報酬 r_1 を投票すると，時間を遡って自チームのプレイヤがボールを保持してから，投票直前の 2 つの連続した自チームプレイヤの行動 $a(t_1)$, $a(t_2)$ と，そのときの局面 $s(t_1)$, $s(t_2)$をエピソード$\{s(t_1), a(t_1), s(t_2), a(t_2)\}$と定義する．また，報酬 r_1 が投票されてから 10 サイクル(1 秒)待って，その間に与えられた報酬 r_2 も加算して学習システムへ引き渡される．

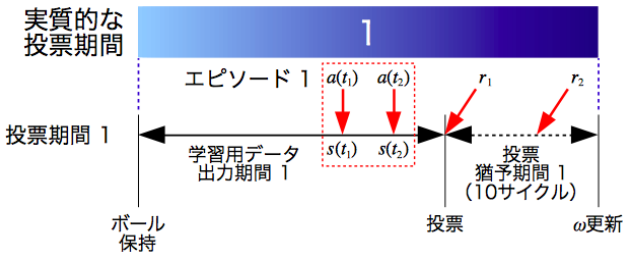

図 3 エピソードと報酬の関係

その他，学習中のチームが得点を挙げた際には報酬を 30 だけ自動的に加算した．また，今回の実験では学習エージェントは同一の評価関数を持ち，ボール保持時の行動決定でのみチェーンアクションを用いると仮定した．

6. 実験

6.1 学習実験

学習エージェントはディフェンダーとゴールキーパーを除く 6 人(FW3 人と MF3 人)とし，評価関数の重み ω の初期値は全て 1 とした． agent2d(ver.3.1.1)を相手に 10 試合行い， 被験者 5 名(A~E)が試合の観戦および投票を行った．チェーンアクションにおける探索深さは 2 段，探索ノード数の上限は 2000 とした．

5 人の被験者のうち，最も学習効果のあった被験者 A の重み ω と報酬 r の推移を図 4 に示す．図 4 では，左の縦軸は重み ω_i (i=1,2,…5)の値，右の縦軸は報酬 r，横軸は学習エピソード回数をそれぞれ表している．ただし，図 4 の報酬 r は 50 エピソードごとに平均化した値である．

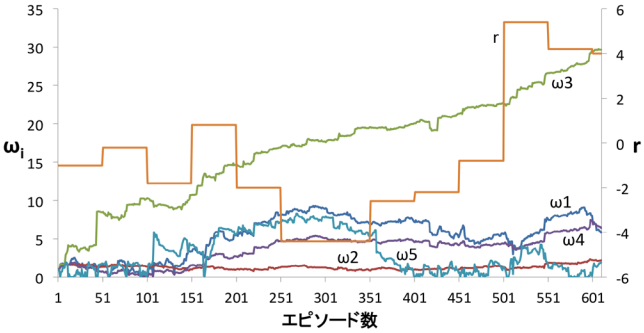

図 4 学習実験における重みと報酬の推移

図 4 では 10 試合終了後には報酬は正の方に大きくなっている．重みは最終的に ω_3 が他と比べて特に大きくなっており，局面評価にボールと両ゴールとの距離が最も重要視されていることがわかる．この特徴は他の 4 人の被験者とも共通して見られ，また，静止画による主観評価値と局面評価関数中の項 $U_i(s)$の値との相関強度を用いた予備実験の結果[8]とも一致している．

6.2 評価実験

未学習のチームと学習後のチームの強さを比較するため，元の agent2d とそれぞれ 300 回対戦を行った．実験条件は基本的に学習実験と同様であるが，評価実験では方策に greedy 選択（ボルツマン分布で T=0 とした場合に相当）を使用した．結果を表 1 に示す．ただし，表 1 における学習後のチームの結果は，最も勝数/負数の値が大きかった被験者 A の重みを用いた実験結果である．なお，勝率の値を計算する際には，引き分けの試合数を取り除いてある．

表 1 被験者 A の評価実験(300 試合)の結果．
ただし，勝率の計算では引き分けを除く．

対agent2d	勝率	引分数	平均 得点 – 失点	ボール 支配率
agent2d	54%	55	2.44 – 2.22	50%
未学習チーム	6%	49	0.18 – 1.63	67%
学習チーム	43%	63	1.30 – 1.57	57%
先行研究[谷川, 2013]	22%	76	0.59 – 1.38	80%

表 1 からわかるように，被験者 A の学習後のチームは未学習のチームと比べて，対 agent2d との対戦において，平均得点が 0.18 から 1.30 と 1.12 点だけ増加したのに対し，平均失点が 1.63 から 1.57 へと 0.06 点だけ減少している．したがって，学習によってチームが強くなったと言える．実際，対 agent 戦における勝率は 6%から 43%へと大幅に向上している．しかし，学習後のチームでも対 agent 戦においての得失点差は，-0.27(=1.30-1.57)とマイナスであり，勝率も 50%にまでは届いていない．

7. スルーパスの出現回数と質に関する考察

評価実験での学習後のチームの対 agent2d 戦の試合内容を観察すると，敵ゴール付近の守備が固まってしまうと攻めることができず，その場で味方とボールのやり取りをするうちに，ボールを奪われたりオフサイドの反則を取られたりする局面が目立った．これが agent2d に対して得点が伸び悩む理由の一つと考えられる．

また，学習後のチームの試合では，得点を挙げた局面の多くがスルーパスの成功によるものであった． 特に，被験者 A はスルーパス実行・成功時に高い報酬を与える傾向があった．そこで，評価実験の試合において出現したスルーパスの回数を集計した．ただし，パスがスルーパスであるかどうかの判定と，それが成功したかどうか，さらにその後の得点につながったかどうかの判定は人間が手作業で行う必要があった．そこで，集計対象は対 agent2d の試合（300 試合）から高得点をあげた上位 20 試合だけに絞ることにした．結果を表 2 に示す．

表 2　評価実験における各チームのスルーパスの出現回数などの集計．ただし，集計対象は 300 試合のうちの各チームが高得点をあげた上位 20 試合に限定．

対 agent2d	スルーパス			スルーパス後シュート成功数
	実行数	成功数	成功率	
agent2d	23	19	83%	10
未学習	44	29	66%	10
学習後	146	121	83%	51

表 2 を見ると，agent2d チームは，回数は 23 回と少ないが味方のレシーバに渡る成功率が 83%と高く，確実に通るスルーパスを出していることがわかる．これに対して，本研究での未学習チームは agent2d チームのほぼ 2 倍の回数のスルーパスを出してはいるが，その成功率は 66%と 17 ポイントも低くなっている．これは 5.2 で述べた探索木の枝刈りの緩和によりスルーパスの候補となるノードが増えたが，安全性を評価する項の重みが適切でなければレシーバが受け取ることができないことを表している．

一方，学習後のチームでは，スルーパスの実行回数は agent2d の 6 倍以上で，かつ，成功率も 83%と高い水準を保っている．しかも，そのスルーパスが通った後に，味方チームがシュートを放って得点を得た回数が agent2d の 5 倍以上である．これは，本研究で行った評価関数の学習が安全性を維持したまま得点に結びつく有効なスルーパスの出現回数の増加に寄与した結果であると言える．

8. まとめ

本研究では，サッカーシミュレーションシステムにおいて，複数のヒューリスティクスで構成された局面評価関数のパラメータの学習に，観戦者の主観評価を報酬として用いる実時間強化学習システムを構築した．本システムを用いて学習実験を行った結果，学習前のチームと比べてスルーパスの出現回数や成功回数が増加し，勝率や平均得点が大幅に増加する可能性があることを確認した．

今回，学習エージェントは同一の局面評価関数を持つと仮定したが，今後は，プレイヤごとの学習や局面評価関数の項の改良，ボールを持っていないプレイヤの行動決定への適用などについて研究を進めて行く予定である．

謝辞 本研究は JSPS 科研費 26330419 の助成を受けた．ここに感謝の意を表する．

参考文献

[1] Itsuki Noda, Hitoshi Matsubara. “Soccer Server and Researches on Multi-Agent Systems,” Proc. of IROS-96 Workshop on RoboCup, pp.1-7 , 1996..

[2] http://wiki.robocup.org/wiki/Soccer_Simulation_League

[3] P. Stone, G. Kuhlmann, M.E. Taylor, and Y. Liu, ” Keepaway Soccer:From Machine Learning Testbed to Benchmark,” in RoboCup 2005:Robot Soccer World Cup IX, eds. A. Bredenfeld, A. Jacoff, I. Noda and Y. Takahashi, pp.93-105, Springer-Verlag, New York, 2006.

[4] S. Kalyanakrishnan, Y. Liu, and P. Stone, ”Half Field Offence in RoboCup Soccer —A Multiagent Reinforcement Learning Case Study,” in RoboCup- 2006:Robot Soccer World Cup X, eds. G. Lakemeyer, E. Sklar, D.G. Sorrenti, T. Takahashi, pp.72-85, Springer-Verlag, 2007.

[5] M. Riedmiller and T. Gabel, ”On Experiences in a Complex and Competitive Gaming Domain -Reinforcement Learning Meets RoboCup-,” Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Games (CIG2007), pp.17-23, 2007.

[6] 秋山英久，”アクション連鎖探索によるオンライン戦術プランニング”，人工知能学会研究会資料，SIG-Challenge- B101-6, pp. 23-28, 2011.

[7] 谷川俊策，五十嵐治一，石原聖司，“RoboCup サッカーシミュレーションリーグ 2D における局面評価関数の設計と学習”，ロボティクス・メカトロニクス講演会 2014，講演番号 1P1-R04.

[8] 田川諒， 谷川俊策， 五十嵐治一，“agent2d のチェーンアクションにおける局面評価関数の重み調整”，FIT 講演論文集，Vol. 13, No. 2, pp.285-288, 2014.

[9] R.J. Williams, ”Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning,” Machine Learning, vol.8, pp.229-256, 1992.

[10] 五十嵐治一， 石原聖司， 木村昌臣， “非マルコフ決定過程における強化学習—特徴的適正度の統計的性質—”， 電子情報通信学会論文誌 D, Vol. J90-D, No. 9, pp.2271-2280, 2007.

[11] 大串明，山本一将，森岡祐一，五十嵐治一，“コンピュータ将棋における方策勾配を用いた局面評価関数の教師付学習”，第 20 回ゲーム・プログラミング・ワークショップ 2015 予稿集，pp.84-87, 2015.

[12] 五十嵐治一，森岡祐一，山本一将，”プロ棋士の棋譜データベースを用いない局面評価関数の学習法についての考察”， 第 34 回ゲーム情報学研究発表会 (2015.7.4，福岡市)，情報処理学会研究報告，Vol. 2015 -GI -34, No. 4, pp.1-8 , 2015.

[13] agent2d のホームページ：<http://rctools.osdn.jp/pukiwiki/index.php?agent2d>