F-032

レシーバの行動決定モデルを利用したスルーパスの強化学習
Reinforcement Learning of Through Passes Using a Receiver's Action Decision Model

山岸 準† 五十嵐 治一†
Jun Yamagishi Harukazu Igarashi

1. はじめに

RoboCup サッカーシミュレーション 2D リーグはソフトウェア同士がコンピュータ上でサッカーをするリーグである。サッカーはチェスや将棋などのボードゲームと異なるいくつかの特徴がある。一つ目はチームプレイが要求されることである。二つ目は、実時間でゲームが行われるので、瞬時に行動を決定しなければならない。三つめは情報が部分的で不確実なことである。サッカーでは自分の視覚内の情報しか取得できず、その情報もノイズを含んでいる。以上の特徴などから、マルチエージェントシステムや協調行動について研究するためのテストベッドとして用いられている [1]。

このリーグでは多くのチームが agent2d [2]というサンプルチームをベースにしている。agent2d(ver.3.0.0)では、プレイヤエージェントが探索木と評価関数を用いてドリブルやパスなどの行動決定を行っている。しかし、agent2d で用いられている評価関数はボールの位置のみを考える単純なものであったため、谷川らが新たに評価関数を作成し、重みの強化学習を行った [3]。しかし、3000 試合学習してもagetn2d に勝ち越すことはできなかった。田川らはこの原因は報酬の質にあると考え、報酬として人間の主観評価を用いるオンライン強化学習システムを開発した。このシステムでは、わずか 10 試合の学習で効果的なスルーパスの発生回数を増加させることができた [4]。

しかし、上記の研究ではボール保持者(パサー)のみが自分の行動を評価し、行動を決定していた。そこで本研究では、パサーにはボール非保持者(レシーバ)の行動評価を考慮する項を、レシーバにはパサーの行動評価を考慮する項を評価関数に導入して、協調行動を促進させる方法を考案した。さらに、田川らのオンライン学習システムに組み込んで学習実験を行ったので、その結果を報告する。

2. サッカーシミュレーション 2D リーグ

RoboCup サッカーシミュレーション 2D リーグは実機を使わず高さがない 2 次元フィールド上で 11 対 11 のプレイヤがサッカーを行うリーグである。このリーグの試合はサーバクライアント方式でシミュレートされており、以下のような流れでシミュレーションが行われている。まず、プレイヤはセンサ情報をサーバプログラム(rcssserver)から取得する。その情報に基づいてプレイヤは各自で行動決定を行い、サーバに kick や dash などの行動コマンドを送信する。この流れを試合終了まで行う。しかし、サーバから受け取るセンサ情報にはノイズが含まれているため、不完全な情報を基に行動決定を行わなければならない。さらに、プレイヤ同士のサーバを介さない直接的な通信はルール上禁止されている。これらの制約があるため、協調行動を実現するための工夫が必要となる。

† 芝浦工業大学 工学部 情報工学科
Shibaura Institute of Technology

3. プレイヤの行動決定

3.1 Chain action について

agent2d には、2010 年から chain action という行動立案の枠組みが導入された [5]。これは、まずパスやドリブルなどボール保持者の行動を「枝」とし、行動後の試合局面(状態)を「ノード」とした探索木を作成する。次に評価関数によって全ノードを評価し、最良優先探索によって評価値の最も大きなノードに至るルート直下の行動が選ばれる。図 1 に chain action の例を示す。この図では、a、b、c が選択対象となる行動、S_1～S_8 が状態、数値がノードの評価値を示している。この例では、S_7 の評価値が最も高いためルートからの次の行動として b が選択される。

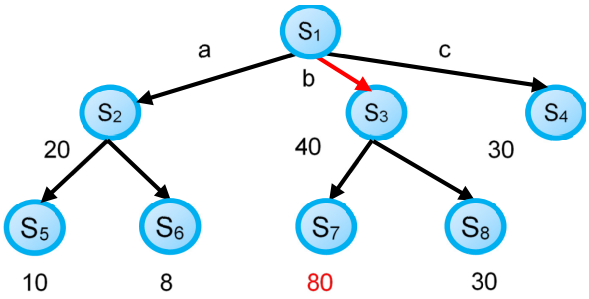

図 1 Chain Action の例

3.2 ボール非保持者への Chain action の適用

agent2d のレシーバの行動決定では chain action を用いず、Delaunay Triangulation を使用してレシーバの移動位置を決定する手法を用いている [2]。しかし、この手法はあらかじめ作成したサンプルを基に移動先の位置を計算する手法であり、敵プレイヤにマークされてもマークを外す動きをしないという問題点がある。そこで大内らはレシーバに

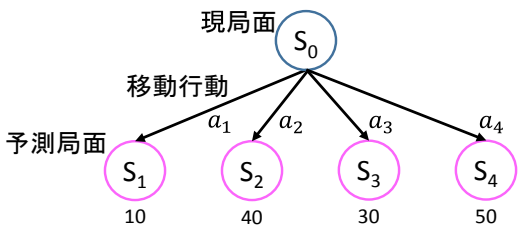

図 2 レシーバの探索木の例 [6]

chain action を適用することを提案した [6]。ただし、計算量の関係で探索木の深さを 1 とした。図 2 にレシーバが作成する探索木の例を示す。

図 2 では、a_1〜a_4 が移動行動、S_1〜S_5 は状態、数値はノードの評価値を表している。この例では、S_5 が最も高い評価値であるため次の移動行動はa_4となる。Chain action の適用と強化学習によりレシーバはパサーにとって良い位置取りをするようになる。その結果、ゴール前でのパス回しによる得点が増加したことが報告されている [6]。本研究ではレシーバの行動選択としてこの方式を利用する。

3.3 確率的方策の適用

agent2d では最良優先探索によって決定論的に行動を決定していた。しかし、谷川 [3]や、田川ら [4]の研究では学習を行うために以下のような Boltzmann 分布による確率的な方策を利用した。

$$\pi(a|s;\omega) \equiv \frac{e^{E_s(S_a;\omega)/T}}{\sum_{x\in A(s)} e^{E_s(S_x;\omega)/T}} \qquad (1)$$

ただし、s_a は局面 s において行動 a 以下の部分木での局面評価値$E_S(s_a;\omega)$が最大の局面（ノード）を表す。また、$A(s)$は局面 s における行動集合、T は温度パラメータ、ω は評価関数中のパラメータである。

4. 評価関数

4.1 本研究で使用した評価関数

agent2d で使用している評価関数はボールと敵ゴール間の距離に依存した単純なものであったが、本研究では複数の特徴を考えた。ただし、パサーとレシーバは同じ形の評価関数を使用する。

$$E_S(s;\omega,\omega') = \sum_{i=1}^{5} \omega_i U_i(s) + \omega_6 U_6(s;\omega') \qquad (2)$$

U_i はそれぞれヒューリスティックスを表した関数であり、ω_i はそれぞれの関数の重みである。U_1 はパス先のパスコースの数、U_2 は agent2d と同様ボールと敵ゴール間の距離、U_3 はボールの x 座標、U_4 は敵ディフェンスラインとの距離、U_5 はシュートコースの広さをそれぞれ評価した項である。U_6 は相互作用の項で次節で説明する。

4.2 相互作用の項 U_6

相互作用の項は味方の行動を考慮する項である。中村ら [7]は 2vs2 のフリーキックにおいて、パサーが予測したレシーバの移動先と自分のパス先との距離の誤差を用いた。

本研究ではフルゲーム(11vs11)において、相互作用の項を導入した。本研究での相互作用の項はパサーとレシーバがお互いに相手の行動がどの程度の評価なのかを評価した項である。すなわち、連携相手の評価関数と重みを用いて自分の行動が相手にとりどのくらいの評価値なのかを計算する項である。具体的には以下の式により計算する。

$$U_6(s';\omega) = 20 \times \frac{1}{1+e^{-E'(s;\omega')/50}} - 10 \qquad (3)$$

$$E'(s;\omega') = \sum_{i=1}^{5} \omega'_i U_i(s) \qquad (4)$$

ここで、ω'は連携相手の重み、E'は連携相手による自分の行動の評価値を表している。

図 3 に例を示す。レシーバがパス先地点として前方を高く評価すると、相互作用の項U_6の値が大きくなり、パサーはレシーバの前方へパスを出す行動を高く評価する。

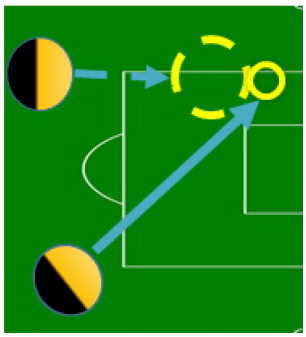

図 3　相互作用の項による効果：パサーからレシーバ.前方へのパス

5. 方策勾配法による重みの学習

5.1 学習則

学習には先行研究[3][4]と同様に方策勾配法 [8] [9]を用いた。まず、学習するエピソードを定義し、エピソード終了後にその時点での局面やエピソード全体を評価して報酬を与える。次に、エピソードあたりの報酬の期待値を極大化するために、確率的勾配法を用いて(2)の各 ω_i を更新する。ここで、方策勾配法による学習則は次のように表される。

$$\Delta\omega_i = \varepsilon \cdot r \cdot \sum_{i=0}^{L-1} e_i(t) \qquad (5)$$

$$e_i \equiv \frac{\partial}{\partial\omega_i} \ln \pi(a(t)|s(t);\omega) \qquad (6)$$

ここで、r は報酬、s(t)は時刻 t における局面、a(t)は時刻 t で選択した行動、L はエピソード長、 ε (>0)は学習係数である。この (6)に(1),(2)を代入した次の式を用いた [4]。

$$e_i = \frac{1}{T}\left[\frac{\partial}{\partial\omega_i} E_S(S_a;\omega) - \sum_x \pi(x|s;\omega) \cdot \frac{\partial}{\partial\omega_i} E_S(S_x;\omega)\right] \qquad (7)$$

$$= \frac{1}{T}\left[U_i(S_a) - \sum_x \pi(x|s;\omega)\cdot U_i(S_x)\right] \qquad (8)$$

5.2 オンライン学習システム

本研究では田川らが作成したオンライン学習システム[4]を使用して報酬を与える。まず、モニターで試合を観測者に見てもらい、投票画面で投票を行う。投票が行われたらコーチは chain action の情報と投票結果を利用してプレイヤの重みを試合中に更新する。このシステムによって、観測者が良いと思った行動が学習とともに増え、最終的に観測者の意図した動きをするようになることが期待される。しかし、人間自身が報酬を与えて学習を行わなければならないため、長時間の集中力が必要になり、試合数が多ければ多いほど学習の質が落ちてしまう危険性がある。そのため、先行研究 [4]ではエピソードを短く定義することにより、どの行動への報酬なのかをある程度限定する。これにより学習時間の短縮を図っている。

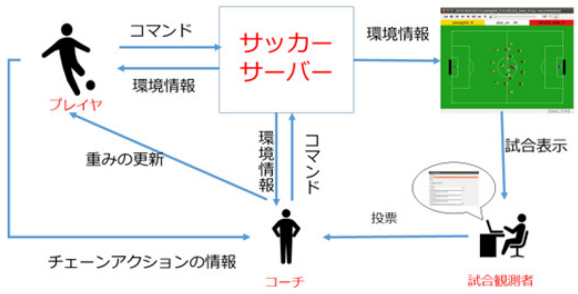

図 4　学習システム

6. 学習実験

学習対象となるプレイヤ数はセンターフォワード(CF)、サイドフォワード(SF)、オフェンシブハーフ(OH)の計 5 名である。実験条件は以下の通りである。

- 対戦相手は agent2d
- 学習試合数は 10 試合
- ボール保持時は全員共通の重み、ボール非保持時はポジションごとに共通の重みを使用
- 重みの初期値はすべて 1
- 温度 T は 80、学習率 ε は 0.1 に設定
- 被験者は 3 名
- 報酬は攻撃中スルーパスや前方にボールが出たときは「good」、攻撃中悪い行動をしたと考えたときは「bad」に投票するように指示

表 1～4 に学習後の重み$\{\omega_i\}$を示す。

表 1　学習後のパサーの重み

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
被験者 1	0.12	1.93	**5.19**	**7.89**	1.04	2.25
被験者 2	**1.93**	0.73	0.63	0.38	**2.21**	1.51
被験者 3	0.38	1.46	**3.61**	**4.06**	0.99	2.27

表 2　レシーバ(CF)の重み

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
被験者 1	0.83	1.02	1.07	1.10	1.00	1.00
被験者 2	0.63	1.04	1.00	1.06	1.00	0.98
被験者 3	**1.27**	0.95	0.86	0.71	0.90	0.97

表 3　レシーバ(SF)の重み

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
被験者 1	0.74	0.98	1.00	**1.20**	1.00	0.95
被験者 2	0.90	**1.20**	**1.20**	1.09	1.00	1.01
被験者 3	1.00	0.87	0.85	0.98	1.00	0.94

表 4　レシーバ(OH)の重み

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
被験者 1	**1.79**	1.09	0.66	0.65	1.00	1.11
被験者 2	1.02	0.95	1.09	1.14	1.00	1.00
被験者 3	**1.63**	1.05	1.08	1.28	1.13	1.09

まずパサーの重みについては、表 1 を見ると被験者 1 と 3 は ω_3 と ω_4 が大きくなった。ω_3 はボールの x 座標、ω_4 はディフェンスラインとの距離に関する項である。従って、前方へのパスやドリブルなどの行動が高く評価される。一方、被験者 2 は ω_1 と ω_5 が大きくなった。ω_1 はパスコースの数に関する項、ω_5 はシュートコースの広さに関する項である。従って、被験者 1 と 3 とは違い、単に前方へパスするのではなく、ボールを受け取った味方にとってパスコースの個数が多い場所にボールを出すようになると考えられる。また、全被験者でパサーにおける相互作用の項の重み ω_6 は増加していた。

次にレシーバの重みについては、表 2 を見ると、被験者 3 の ω_1 が高くなった。従って、被験者 3 の CF はパスコース確保に向かうと考えられる。

表 3 を見ると、被験者 1 は ω_4、被験者 2 は ω_2 と ω_3 が比較的大きくなっている。これらの項はいずれもボールの位置座標に関する項なのでレシーバは前方に移動しやすい。

表 4 を見ると、被験者 1 と 3 の ω_1 が大きくなっている。従って、被験者 1 と 3 の OH はパスコースを確保するような移動をすると考えられる。

7. 評価実験

7.1 実験内容

学習チームのスルーパスの実行回数と強さを調べるために、オリジナルの agent2d (ver.3.1.1)と 500 試合対戦させた。また、このうちの得点が多かった 10試合を選んで、スルーパスの実行回数を目視で計測した。ただし、「スルーパス」とはディフェンスラインの裏へ出したパスと定義する。結果を次節で述べるが、勝率の計算では引き分けを除いてある。

7.2 勝率とスルーパスの回数

表 5 と表 6 に agent2d との対戦結果とスルーパスの実行回数に対する計測結果(10 試合)を示す。

表 5　agent2d との対戦結果

	勝-負-分	勝率[%]	得点	失点
学習前	161-237-102	40.5	2.00	2.39
被験者 1	217-180-103	54.7	2.59	2.37
被験者 2	151-244-105	38.2	1.78	2.29
被験者 3	266-137-97	**66.0**	2.72	2.08

表 6　スルーパスの計測結果

	実行数	成功数	成功率[%]
学習前	25	22	88.0
被験者 1	13	11	84.6
被験者 2	**44**	**36**	81.8
被験者 3	18	16	88.9

表 5 より、被験者 1 と 3 では学習前より勝率が上がり、agent2d に勝ち越したことがわかる。これは、平均得点が両方とも 0.5 ポイント以上増加していることから、得点能力が向上したものによると考えられる。しかし、表 6 より被験者 1 と 3 はスルーパスの実行回数は増加していない。これはスルーパスはなかなか生成されない行動であるため、学習中にあまり出現しなかったためだと考えられる。

一方、被験者 2 は表 6 からスルーパスの実行回数の増加に成功していることがわかる。しかし、スルーパスの増加が勝率の向上には貢献せず、勝率が学習前とほとんど変わっていないことが表 5 からわかる。

7.3 得点シーンの分析

前節で述べたように、被験者 1、3と 2 ではスルーパスの実行回数と得点能力が異なっていた。この原因を調べるために、試合を観測し、得点直前のシーンにおける攻撃側の行動パターンを次の 3 つに分類することにした。

- スルーパス：スルーパスを出した後ドリブルからシュートした場合
- パス回し：ペナルティエリア内でパスを回してシュートをした場合
- ドリブル：ペナルティエリア外からドリブルして直接シュートした場合。スルーパスからのシュートは除く

計測に用いた試合は 7.2 の実験と同じく、agent2d と対戦させて高得点を上げた 10 試合である。結果を表 7 に示す。

表 7　得点シーンにおける行動分析[%]

	スルーパス	パス回し	ドリブル	計
学習前	**20.8**	66.0	**13.2**	100
被験者 1	6.4	**92.1**	1.6	100
被験者 2	**36.7**	61.2	2.0	100
被験者 3	4.2	**94.4**	1.4	100

表 7 より被験者全員において、ドリブルでの単独プレイからの得点回数が減少し、味方とスルーパスを通したり、パスを回して得点する行動が増加していることがわかる。特に、被験者 2 のチームはスルーパスからの得点が増加し、被験者 1 と 3 はゴール前でのパス回しから得点する回数が大幅に増加していた。表 5 で見られた勝率の向上は、このゴール前のパス回しがうまくなり得点力が向上したためと考えられる。このようにゴール前のパス回しがうまくなったのは、表 2や表 4から分かるように、レシーバ(CF や OH)がパスコースの個数が多い地点へ移動するように重みを学習したためと考えることができる。

8. おわりに

本研究では、サッカーシミュレーション 2D で行動の評価関数に味方の行動評価を組み込んで協調行動を促進させることを試みた。さらに、人間の主観評価による報酬を用いた方策勾配法により評価関数の重みを学習した。その結果、10 試合の対戦で、ある被験者ではスルーパスの実行回数が増加し、別の被験者ではゴール前のパス回しが強化された。その結果、agent2dの対戦勝率を学習前の 40.5%から最高勝率 66.0%まで強化することができた。

今後の課題の一つ目として、報酬の自動化が考えられる。現在、報酬は人間が試合を観測し、そのプレイが良いか悪いかを判断して与えている。そのため、長時間の学習が困難であり、試合数が限られている。従って、報酬を自動的にうまく与えられるシステムが作成できれば、より多くの試合を行い方策勾配法によりさらなる強化学習ができるのではないかと考えている。また、課題の二つ目として、教師あり学習の導入が考えられる。現在は確率的方策による強化学習によって学習を行っているため観測者が行ってほしい行動をしない場面が多々ある。従って、観測者自身が行ってほしい行動を正解行動として与える教師あり学習を導入することにより効率よくプレイヤの行動が学習できるのではないかと考えている。

参考文献

[1] Noda Itsuki, Matsubara Hitoshi, "Soccer server and researches on multi-agent systems." Proceedings of the IROS-96 Workshop on RoboCup, pp.1-7, 1996.
[2] Akiyama Hidehisa, Nakashima Tomoharu, "Helios base: An open source package for the robocup soccer 2d simulation." RoboCup 2013:Robo World Cup XVII, pp.528-535, 2013.
[3] 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学習." ゲームプログラミングワークショップ 2013 論文集, pp.106-109, 2013.
[4] 田川諒,五十嵐治一, “サッカーエージェントにおけるスルーパスの強化学習”. 第 15 回情報科学技術フォーラム(FIT2016), pp.267-272, 2016.
[5] 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.
[6] 大内斉,五十嵐治一, "局面評価関数を用いたサッカーエージェントの移動先決定." ゲームプログラミングワークショップ 2016 論文集, pp.49-56, 2016.
[7] 中村浩二,五十嵐治一,石原聖司, “方策勾配法を用いたサッカーエージェントの学習～フリーキックにおけるキッカーとレシーバ”, 第 23 回 SIG-challenge 研究会予稿集, pp.7-12, 2006.
[8] Williams, Ronald J, "Simple statistical gradient-following algorithms for connectionist reinforcement learning." Machine learning, Vol.8, pp.229-256, 1992.
[9] 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.