方策勾配法を用いたサッカーエージェントの学習：パス・レシーブ

五十嵐治一† 福岡仁志† 佐野直人† 石原聖司‡
芝浦工業大学工学部情報工学科† 近畿大学工学部電子情報工学科‡

1. はじめに

RoboCup サッカーシミュレーション 2D リーグでは，実環境下でロボットを動かす難しさから解放されるため，複数エージェントによる協調的な行動に研究の焦点を当てることができる．しかし，マルチエージェント学習には，状態空間の爆発問題，同時学習問題，不完全知覚問題，報酬割り当て問題等のマルチエージェント系特有の難しさがあるとされている[1]．これら四つの問題に対して注意を払いながら．本研究では，フルゲームにおいてプレーヤが味方へパスを出す際のパス先の選択問題を学習対象とした．

2. パス選択問題

本研究で取り扱うパスは，「ダイレクトパス」と「スルーパス」との 2 種類である．前者は，味方プレーヤが位置する地点へのパスであり，ボールを自チーム内で支配し，試合を優位に進めるためのパスである．後者は，敵の守備ラインの裏のスペースへ出す攻撃的なパスである．本研究では，この 2 種類のパスのために，パスを構成する 2 人のプレーヤ，passer と receiver に表1に示す行動決定を学習させた．表1で，ダイレクトパスにおける receiver の学習対象は，もし，passer と自分との間のライン（=パスコース）上に敵が存在する場合に，レシーブしやすくするために移動する移動先の地点（レシーバの周囲を格子状に分割）を決定する方策である．また，スルーパスにおける receiver は，最適と思われるパス先の地点（敵陣を格子状に分割）決定し，passer へパスを要求する．passer はそれに従うのでこの際には特別な学習はしない．passer は2種類のパスを切替える必要があるが，その判断は今回は固定の方策を用いて学習対象外とした．

表1　2種類のパスにおける学習対象

	ダイレクトパス	スルーパス
passer	receiver の選択	-
receiver	パスコース上に敵がいた時の自分の移動先地点（セル）の選択	自分の移動先地点（セル）の選択

"Learning of Soccer Player Agents Using a Policy Gradient Method: Pass Selection and Pass Receive"
†Harukazu IGARASHI, Hitoshi FUKUOKA, Naoto SANO, Shibaura Institute of Technology
‡Seiji ISHIHARA, Kinki University

3. 方策勾配法によるマルチエージェント学習

3.1 方策勾配法

エージェントの学習には強化学習の一種である方策勾配法[2]を用いる．方策勾配法は，報酬の期待値が最大（極大）になるように方策中のパラメータを更新する学習法である．このときの最大化の手段として確率的勾配法を用いる．方策として if-then 型のルールや，ポテンシャルなどの様々な関数が利用でき[3]，環境や方策に関するマルコフ性を要求しないで用いることも可能なので[4]，方策における知識表現が容易であるという長所がある．

3.2 学習則

一般に，プレーヤλが状態（=λの周囲の局面）s_λ において行動 a_λ を決定する問題を考える．具体的には，passer が行うパス先の相手や地点の決定，receiver が行う移動先地点（セル）の決定の問題である．今，行動 a_λ を評価するのに有用なヒューリスティクスを $U_j(a;s)$ $(j=1,2,\cdots,N_U)$ を用意し，これらの線形和で表現される次の目的関数 $E(a_\lambda;\ s_\lambda,\{\omega_j^\lambda\})$ を考える．

$$E_\lambda\left(a_\lambda;s_\lambda,\left\{\omega_j^\lambda\right\}\right)=-\sum_j \omega_j^\lambda \cdot U_j\left(a_\lambda;s_\lambda\right) \qquad (1)$$

具体的なヒューリスティクスの形は問題ごとに異なるが，$0\leqq U_j\leqq U_{max}$ というようにある程度正規化された関数として設計する．重み係数については $\omega_j^\lambda\geqq 0$ と仮定する．

上で定義した目的関数を用いて，プレーヤλの方策を次の Boltzmann 分布関数で与える．

$$\pi_\lambda\left(a_\lambda;s_\lambda\right)\equiv\frac{e^{-E_\lambda(a_\lambda;s_\lambda)/T_\lambda}}{\sum_a e^{-E_\lambda(a;s_\lambda)/T_\lambda}} \qquad (2)$$

このとき，(1)の重み係数は次の学習則に従い，エピソードσの終了時に更新される[3]．

$$\Delta\omega_j^\lambda=\varepsilon\cdot r(\sigma)\sum_{t=0}^{L(\sigma)-1} e_{\omega_j^\lambda}(t) \qquad (3)$$

ただし，

$$e_{\omega_j^\lambda}(t)=\frac{1}{T}\left[U_j\left(a_\lambda(t)\right)-\sum_{a_\lambda}U_j\left(a_\lambda\right)\pi_\lambda\left(a_\lambda;s_\lambda(t),\left\{\omega_j^\lambda\right\}\right)\right]$$

であり，$L(\sigma)$ はエピソード長，$r(\sigma)$ はそのエピソードを評価して与えられた報酬である．

4. ダイレクトパスの学習

4.1 報酬

自チームがボールを保持してから敵チームにボールを奪われるまでの状態・行動列をエピソードとする．報酬 r はエピソード長 L の逆数の逆符号，$r \equiv -1/L$ (<0) とする．したがって，自チームがボールを得てからの保持時間が短いほど大きな罰が，passer と receiver に与えられる．

4.2 ヒューリスティクス

passer の学習においては，(i)パスコース上に敵がいなければパスは成功しやすい，(ii)パス先と敵とが離れているほど成功しやすい，(iii)パス先周辺に敵が少ない方が成功しやすい，(iv)パス先がゴールに近いほど攻撃的でパスとしての価値は高い，(v)パス先の位置情報が最新であれば信頼性が高く成功しやすい，という 5 種類のヒューリスティクスを用いた．

receiver の学習では，上記の(i)～(iv)において「パス先」を receiver の「移動先」と置き換えた関数をそのまま使用した．さらに，(v)移動距離が短い方が望ましい，という関数を加えた 5 種類のヒューリスティクスを用いた．

4.3 学習実験と評価実験

世界的に広く用いられている UvA Trilearn Base 2003 チームをベースに，基本的な技や戦略を加えた上に，Midfielder(MF) 4 名にダイレクトパスの学習機能を移植し，未学習チーム（重み ω_j^{λ} が全てゼロなので MF はランダムにパス相手や移動先を選択する）と 50 試合対戦させる学習実験を行った．評価実験として，T を極めて小さい値に設定（決定論的方策）し，未学習チームとの対戦を 30 試合行った．その結果を表 2 に示す．MF によるボール支配力が高まり，失点が減少し，勝利数が増加したことがわかる．

表 2　ダイレクトパスの学習：評価実験結果

学習試合数	勝敗	得点-失点
0	15 勝 13 敗 2 引分	24-25
50	21 勝 9 敗	27-14

5. スルーパスの学習

5.1 報酬

まず，レシーバがパスを要求し，passer がそれに応じてから，敵にボールを奪われるかプレーが途切れるまでを 1 エピソードと定義する．

報酬 r は 3 つの報酬 r_1, r_2, r_3 の和で与えた．r_1 は receiver がレシーブできた地点が敵のオフサイドラインよりも敵ゴール側であれば，その距離が長いほど大きくなる関数である．r_2 は上記地点と敵ゴールとの距離が長いほど大きくなる関数である．r_3 は固定値で，レシーブ後にそのエピソード内で自チームが得点できたときに与える．

5.2 ヒューリスティクス

receiver の学習においては，4.2 の(i),(ii)と,(iii) receiver とボールのレシーブ地点までの到達時間が拮抗しているほど攻撃的である，(iv)receiver の到達時間がボールのそれよりも早い必要性がある，(v)レシーブ地点が敵のオフサイドラインよりも敵ゴール側であるほど攻撃的である，(vi)レシーブ地点が敵ゴールに近いほど攻撃的でより望ましい，という 6 種類のヒューリスティクスを用いた．

5.3 学習実験と評価実験

日本で広く用いられている Helios(agent2d ver.1.0.0[5])チームの左右の forward (2 名) にスルーパスのレシーバとしての学習機能を組み込んだ．学習機能を組み込んでいないチームと 50 試合対戦させて学習させた後，評価用に 30 試合対戦させた．評価実験の結果を表 2 に示す．表 2 中の「パス成功回数」とは，オフサイドラインよりも敵ゴール側でレシーブ出来たスルーパスの成功回数（30 試合中）である．スルーパスの成功回数が学習前（重み ω_j^{λ} がすべて 0）の 3 倍になり，得点も 2 点増えている．6 得点のうち 5 点はスルーパスの成功による得点である．実際の試合内容を観察しても，学習後は攻撃的で有利に試合を進めている場合が多く学習効果がよくわかる．

表 3　スルーパスの学習：評価実験結果

チーム	勝-負-引分け	得点-失点	パス成功回数
学習前	4-0-26	4-0	6
学習後	6-2-22	6-3	17

参考文献

[1] 高玉圭樹：マルチエージェント学習－相互作用の謎に迫る－，コロナ社(2003)．

[2] Williams,R.J.: Simple Statistical Gra-dient Following Algorithms for Connectionist Reinforcement Learning, Machine Learning, vol.8, pp.229-256(1992).

[3] 石原聖司，五十嵐治一：マルチエージェント系における行動学習への方策こう配法の適用-追跡問題-，電子情報通信学会論文誌 D-I, Vol.J87-D1, No.3, pp.390-397(2004).

[4] 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).

[5] http://rctools.sourceforge.jp/pukiwiki/