方策勾配法を用いたサッカーエージェントの学習
～フリーキック時の壁パスとゲーム中のパス選択～

○五十嵐 治一 (芝浦工業大学工学部)　中村 浩二　(芝浦工業大学工学部)
福岡 仁志　(芝浦工業大学工学部)　石原 聖司　(近畿大学工学部)

本研究は複数のエージェントによる協調行動の学習法の開発を目的としている．その題材としてRoboCupサッカーシミュレーションリーグにおけるゴール前でのフリーキックの問題と，フルゲーム中でのボールキープ問題とを取り上げた．行動決定にヒューリスティクスを用いるために，方策における知識表現が容易である方策勾配法を学習法として用いた．実験結果は2対2における壁パスが実現され，パスやドリブルによるボールキープに対して本方式が有効であることを示している．

Learning of Soccer Player Agents Using Policy Gradient Method
~ Wall Pass after Free Kicks and Pass Selection in a Full Game ~

* Harukazu IGARASHI(SIT), Koji NAKAMURA(SIT), Hitoshi FUKUOKA(SIT),
Seji ISHIHARA(Kinki Univ.)

This research developed a learning method for the coordination of multi-agents. We dealt with two problems in RoboCup Soccer Simulation games. The first problem is free kicks in front of the opponent goal. The second is pass selection during a game. The policy gradient method is applied as a learning method to solve the two problems because it can easily represent various heuristics for pass selection and pass receiving in a policy function. Experimental results show that our method effectively realizes wall passes after free kicks in 2 v 2 mini-games and clever pass selection of the four midfielders in a full game.

1. 研究背景と目的

近年，人工知能の学習の分野ではマルチエージェント環境下での協調行動，実時間処理，不完全知覚における学習といった複雑な問題が研究されている．これらの問題を含んだ題材としてロボットサッカーの競技会であるRoboCupが提唱されている[1][2]．これまでに著者らは，ゴール前でのフリーキックの場面においてpasserとreceiverによる協調行動の学習という研究を行ってきた[3]．しかし，この研究では２人のプレーヤー（passerとreceiver）が行動決定を同時刻に各々１回のみ(passerのパス先決定とreceiverの移動先決定)行うというものに留まっていた．

そこで本研究では，フリーキック時にパス行動を複数回行うという協調行動を目指した．すなわち，フリーキックからの壁パスの実現である．これを研究課題１とする．次に，同じ手法をフルゲーム中でのボールキープ問題(Keepaway) [4]へ適用することを試みた．これを研究課題２とする．

2. サッカーシミュレータ

本研究はシミュレーション環境としてRoboCupの公式シミュレータであるSoccer Server[2]を用いる．ユーザはクライアントとして11人分のプレーヤープログラムを作成し，サーバに接続する．クライアント同士は直接通信を行うことはできないため，チームの制御は独立したクライアントによる自律分散型制御である必要がある．

3. 学習方式

3.1 行動決定方式

本研究では，以下のような行動決定方式を用いる[3]．行動 a_k の価値を次のような目的関数で表す．

$$E(a_k;\omega) = -\sum_i \omega_i \cdot U_i(a_k) \qquad (1)$$

この関数は，行動 a_k の評価の上で有効と思われる特徴量（ヒューリスティクス）U_i の線形和である．ただし，目的関数 E の値が小さい方が行動としての価値が高くなるように設計する．また，重み ω_i は学習により決定する．

この目的関数を用いて，プレーヤーは次のボルツマン選択による確率的な方策 $\pi(a_k;s)$を用いて行動 a_k を決定する．

$$\pi(a_k;s) \equiv \frac{e^{-E(a_k;s)/T}}{\sum_{a_k \in A} e^{-E(a_k;s)/T}} \qquad (2)$$

なお, s は全系の状態(i.e. 全プレーヤーとボールの位置),T は決定のランダム性を表すパラメータである．

3.2 学習則

複数回の行動決定を含むあるエピソードを定義し，エピソード終了時に結果に対して報酬を与え，(1)の ω_i を方策勾配法[4]の次の更新則[3]により学習する．

$$\Delta\omega_i = \varepsilon \cdot r \cdot \frac{1}{T} \cdot \left(U_i(a_k) - \sum_a U_i(a) \cdot \pi(a,s) \right) \qquad (3)$$

ここで ε は学習係数，r は報酬である．なお，本研究ではエージェントごとに，かつ，行動決定ごとにそれぞれ目的関数と報酬とを用意した．

4. 研究課題１：壁パス

4.1 フィールドの分割

実験では図１に示すようにペナルティエリア付近を 35 個のセルに分割したフィールドを用いる．プレーヤーは３種類で，味方プレーヤーが２人，敵チームのディフェンダー１人とゴールキーパー１人である．また，便宜上味方プレーヤーは最初にパスを行うものをA，最初にレシーブを行うものをＢとする．

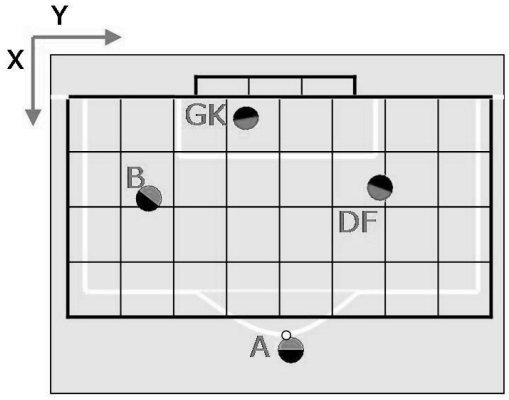

Fig. 1 Arrangement of players

4.2 処理の流れ

エピソード中の処理の流れを図２に示す．なお，行動決定の機会をA1~A4, B1~B2で表す．まず, passerであるプレーヤーAはフリーキック前にパス先を決定する(A1)．フリーキック後，Aは移動先を決定後(A2：ポジショニング)，そこへ移動し，決定した２つの行動内容（A1とA2）をBへ通知する．receiverであるBはその情報に基づいて移動して(B1)ボールを受け取った後はpasserとなり，逆にAはreceiverへと役割を変更する．次に，新たにpasser となったBはパス先を決定して(B2)パスを行い，A に行動内容を通知する．Aはその情報を基に移動しボールを受け取る(A3)．パスを受けたA は必ずシュートを行うものとする(A4). 今回，これらの行動決定のうちでA1, A2, B2を学習の対象とした．また，パス先と移動先は図１のように35個のセルで離散的に表現されている．

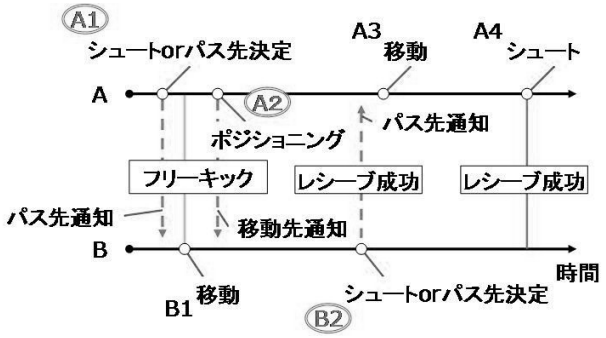

Fig.2 Actions of players A and B

なお，１エピソードを70シミュレーションサイクル(7sec)とし，１エピソード終了前にゴールを決めるか70シミュレーションサイクルが過ぎると，エピソードはそこで終了とし，報酬を与える．

4.3 実験条件

複数回の行動に対して学習をする前準備として，A1 のみを学習する予備的な学習実験を行った．この結果を本実験において，A1 の重みの初期値として使用することで，学習時間の短縮ができる．予備実験の際のパラメータは T=10.0, ε=0.04 とし 2000 エピソード学習する．

次に本実験を行う．ここで学習が期待されるのは A→B→A とパスが渡り A4 でシュートを行うまでの，いわゆる壁パスが通った後にシュートを行う協調プレーである．実験の際のパラメータは T=10.0, ε=0.04 とし，2000 エピソード学習する．なお，予備実験，本実験共に対戦相手として Trilearn Base を使用した[5]．Trilearn Base とは 2003 年に優勝したアムステルダム大学チーム UvA Trilearn 2003 から高度な行動決定や戦略を除いてソースコード形式で配布されているチームである．よって，UvA Trilearn 2003 よりは弱くなっているが最低限の行動は保証されている．

4.4 ヒューリスティクスの設定

本研究では行動 A1,B2 におけるパス先の決定において,ヒューリスティクスとして次の U_1~U_4 を用いた：U_1=パスコースにおける敵の有無，U_2=パス先とゴールとの距離，U_3=パス先と最も近い味方との距離，U_4=パス先と最も近い敵との距離. なお,B2 で U_3 を用いる際には，A は A2 で B に通知した移動先に行くことを前提としている．

また，行動 A2 における移動先(ポジショニングの位置)の決定において,ヒューリスティクスとして次の U_5~U_{10} を用いた：U_5=移動先とパス先との距離，U_6=移動先とゴールとの距離，U_7=移動先と最も近い味方との距離，U_8=移動先と最も近い敵との距離，U_9=移動先と現在の自身との距離，U_{10}=移動先とパス先の間の敵の有無．これを(1)へ代入し，行動を決定する．なお，それぞれの重み ω_i の初期値は 10 から 15 の間でランダムに設定した．

4.5 報酬

4.3 の学習では，行動決定 A1，A2，B2 ごとに目的関数を用意する．プレーヤーA は A1，A2 の，プレーヤーB は B2 の目的関数を各行動決定に用いる．したがって，重み係数も別々に{ω_{A1}}，{ω_{A2}}，{ω_{B2}}と用意する．また，それぞれの重み係数の学習に対し，学習則(3)において与える報酬 r_{A1}，r_{B2}，r_{A2} を図３のように設定し，1 エピソード終了時に与える．

それぞれの重みに与えられる報酬のパターンとしては 6 種類あるが，行動決定の段階(パスの図２における A1~B2，B2~A4，A4 以降の三種)と結果に応じていずれか一種類の報酬が与えられる．

	r_{A1}	r_{B2}	r_{A2}
A1でのパス失敗	−20		
直接シュート成功	60		

A1のパス成功

	r_{A1}	r_{B2}	r_{A2}
B2でのパス失敗	0.5	−20	−20
B2でシュート成功	80	80	

B2のパス成功

	r_{A1}	r_{B2}	r_{A2}
A4でのシュート失敗	2	2	2
A4でシュート成功	100	100	100

Fig.3 Rewards r_{A1}, r_{B2}, and r_{A2}

4.6 学習実験の結果

4.3 の条件で学習実験を行った．100 エピソードごとの報酬の期待値 $\overline{r_{A1}}$, $\overline{r_{B2}}$, $\overline{r_{A2}}$ を図４に示す．また，表１に重み値の学習結果を示す．

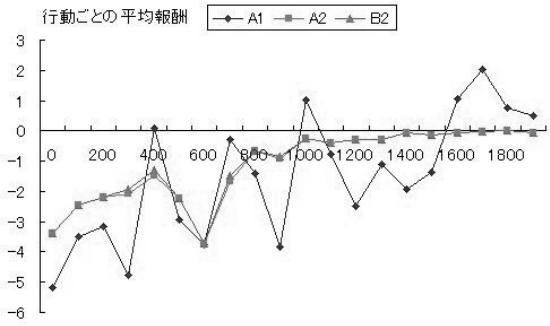

Fig.4 Expected value of rewards $\overline{r_{A1}}$, $\overline{r_{A2}}$, and $\overline{r_{B2}}$

Table1 Change of weights $\{\omega_{A1}\}$,$\{\omega_{A2}\}$,and $\{\omega_{B2}\}$

(a)Action A1

	ω_1	ω_2	ω_3	ω_4
学習前	13.02	4.52	23.09	16.12
学習後	0.85	0.04	27.75	13.74

(b)Action A2

	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}
学習前	10.74	13.11	10.85	14.32	14.57	12.30
学習後	2.73	2.80	2.97	23.53	18.69	14.79

(c)Action B2

	ω_1	ω_2	ω_3	ω_4
学習前	12.07	13.35	12.82	13.11
学習後	10.69	4.48	18.91	15.63

まず，表１a より，行動決定 A１では学習後の重み ω_3 と重み ω_4 の割合が大きい．したがって，最初に A が行うパス(A1)では味方の近く(U_3)で敵から遠い場所(U_4)，つまり B が安全にパスを受けることのできる場所へ蹴ることを学習している．

同様にして，表１b より，行動決定 A2 で移動先を決定（ポジショニング）を行う先を決定する際に，学習後の重み ω_8 と重み ω_9，そして重み ω_{10} の割合が大きい．したがって，敵から遠く(U_8)，パスコースに敵がいない(U_{10})，がつ，自分から近い場所(U_9)．つまりパスカットの行われにくい近くの場所へ移動することを学習している．

表１c からは，行動決定 B2 では ω_3 と ω_4 が大きくなっており，A1 と同様に B は A が安全にパスを受けることのできる場所へパスすることを学習していることがわかる．

5. 研究課題２：ボールキープ

5.1 学習対象

本章では，フルゲーム時のドリブルとパス先選択の学習に，３章で述べた学習方式を適用する．学習対象はボールを保持しているプレーヤ（ボールホルダー）のみとし，レシーブ行動の学習は行なっていない．また，ミッドフィルダー(MF)がパスとドリブルを最も多用するのでMF４人のみが学習を行う．

5.2 行動集合

ボールホルダーの行動集合はゴールキーパーを除く味方の選手９名へのパスとドリブルとする．「ドリブルをする」は自分自身へのパス行動と表現される．ボールホルダーは式(1)による確率的な方策を用いてレシーバkを決定する．なお，ヒューリスティクスには4.4のU_1，U_2，U_4を用いた．

5.3 報酬関数

本実験では，マイボールになってからボールを奪われるまでを1 エピソードとする．荒井らは３対２のKeepaway問題において[6]，Sarsa法適用時のベストな報酬関数を実験的に求めている[7]．今回，荒井らと同様に，エピソード終了時に報酬$r(t_j)=-1/f(t_j)$を与える．ただし，$f(t_j)$ はエピソード終了時刻と各学習エージェントj の最終行動の時刻との差である．この報酬関数では，ボールを奪われる直前の行動ほど罰が大きくなっている．

5.4 使用するエージェント

本実験では11対11のフルメンバーでの試合を扱う．学習を行う味方エージェントのチームには，Trilearn Base 2003をベースにして我々が開発したチームを用いた．元のTrilearn Base 2003チームは，視覚情報の管理が弱い上，インターセプト，タックル，クリア，マークなどの防御機能がなく，パスやドリブルなどもない．今回は手動で視覚情報の管理の強化，防御機能の追加と，FWにはシュート力の向上を行い，さらにDF, FWにも簡単なパスやドリブルの機能の追加など，個人技を強化した．対戦チームには電気通信大学が開発したYowAI 2003を用いた．YowAI 2003は’03年RoboCupジャパンオープンで優勝し，同年RoboCup世界大会で2次リーグ進出を果たしたチームである．スタミナ管理に定評があり，パスを主体とした攻撃をする．

5.5 実験結果

以上の条件の下で，YowAI 2003を相手にして50試合学習を行った．評価実験用として，学習後と未学習のチームをそれぞれYowAI2003と50試合対戦させ，対戦の結果をSoccerScope2003[8]により解析した．諸量の一試合あたりの平均値を表2 に示す．なお，学習時の重み$\{\omega_i\}$の初期値は0.0，$\varepsilon = 0.1$, $T = 10.0$とし，評価実験ではT=0.1の，ほぼ決定論的な方策を用いた．

Table2 Stats changes before Learning and after Learning

	ゴール数	失点数	シュート数	ドリブル数
学習前	3.74	0.32	8.44	26.88
学習後	4.78	0.16	11.24	46.14

	パス本数	パス成功率	レシーブ位置の平均×座標（全体）	（前衛三人）
学習前	86.34	68.7%	-5.3708	7.888
学習後	66.02	72.0%	0.6926	13.1988

表２から，学習後は学習前と比べてシュート数と得点数が増えており，チームが攻撃的，かつ，強くなっている．この原因を分析すると，パスの本数が約20 本減っているが，その代わりにドリブルの回数が約20 回増えている．これは無理なパスを出して奪われるよりは，ドリブルを多用する戦術に切り替えたと考えられる．パスの成功率の上昇は3ポイントにしか過ぎないが，それはボールを受ける位置が5~6m敵陣寄りになり，パスがより難しい場面が増えたためである．ボールの軌跡を解析した例(一試合分)を図５に示す．これからも，学習後は，敵陣によく攻め入っているのがわかる．敵陣内でのプレーが増えたことが，シュート数と得点数の向上に寄与している．

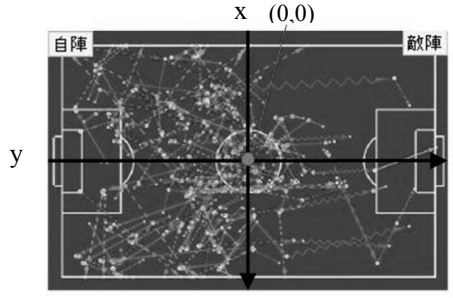

Fig.5(a) Ball Position before Learning

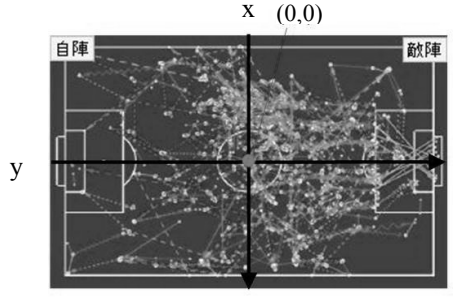

Fig.5(b) Ball Position after Learning

また，学習後に実際の試合の様子を観察すると，MF がドリブルで相手のディフェンスラインに突進し，FWをマークしていた敵DF がつり出されてFWがフリーになった瞬間にパスを出す，といった高度な行動が見受けられた．これは，明示的に「このような作戦を取れ」と教示したわけではなく，本手法によりエージェントが局面によってパスとドリブルを使い分けることに成功したと言える．

6. 今後の課題

研究課題１では，複数パス交換へと協調行動を拡張することを目的とし，壁パスの実現に関して学習の有効性を確認することができた．しかし，安全なパスを目指すあまりに得点率の向上までは達成できなかった[9]．得点率の向上をさせるためには，報酬の設計を再検討する必要がある．たとえば，バックパスを行う場合は，パス成功に対する報酬を減少させ，シュートを成功させる方がそれよりももっと高い報酬を与えることなどが考えられる．しかし，必ずしもバックパスが悪いとは限らないため，報酬の設計は慎重にしなければならないであろう．

研究課題２では MF のパス選択というタスクの学習を取り扱ったが，学習で得られた方策をそのまま DF や FW に持たせても，それだけではチームは強くならない．パス選択以外にも，DF にはボールクリアというタスクが, FW には精度の高いシュートをするというタスクがあり，それぞれの方策を学習する必要がある．さらに，これらのタスクにおいて適切な方策を学習できたとしても，パス選択の方策とこれらの行動の方策とを切り替えて用いるタイミングなどの上位レベルの方策が必要である．このような階層的な方策学習も本方式の強化学習で実現可能かどうか研究を進めて行きたい．

参考文献

[1] 野田五十樹：“シミュレーションリーグとインフラ技術の技術的課題と展望”，日本ロボット学会誌，Vol.20,No1,pp.7-10(2001)

[2] http://sserver.sourceforge.net

[3] 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)

[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, Vol.J90-D, No.9，pp.2271-2280(2007)

[5] The Universiteit van Amsterdam 〈http://staff.science.uva.nl/~jellekok/robocup/〉

[6] Stone,P.，Sutton,R.S.，and Kuhlmann,G.：“Reinforecement Learning for RoboCup Soccer Keepaway”，Adaptive Behavior，Vol.13，No.3，pp.165-188(2005)

[7] 荒井幸代, 田中信行: “マルチエージェント連続タスクにおける報酬設計の実験的考察－RoboCup Soccer Keepaway タスクを例として－”，人工知能学会論文誌, Vol. 21, No. 6, pp.537–546 (2006).

[8] http://ne.cs.uec.ac.jp/%7Ekoji/

[9] 福岡仁志，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～フリーキック時の壁パス～”，第25回 SIG-Challenge研究会予稿集，pp.74-77((2007)