F-034

サッカーシミュレータ環境における強化学習を用いたパスの学習
Learning to pass the ball Using Reinforcement Learning on the Soccer Simulator

高橋 昌生 † 岡 夏樹 †
Masaki Takahashi Natsuki Oka

1. まえがき

近年，マルチエージェントシステムに関する研究が注目を浴びており，複数のエージェントが互いに協調することで問題を解決する協調行動の実現は，工学的及び認知科学的観点から非常に興味深い話題である．しかし，複数のエージェントの存在する環境におけるすべての行動を設計するのは非常に困難であり，学習によるエージェントの行動獲得手法の研究に対する期待が高まっている．学習手法のひとつに，強化学習がある．これは，試行錯誤を通じて未知の環境に適応する，という学習の枠組みであり，人間が正しい解を教えてやらなくともエージェント自身が解を見つけることが出来るため，エージェントの行動学習の手法として期待されている．本研究では，マルチエージェントシステムの標準問題の 1 つとして多くの研究がなされている RoboCup Soccer Simulator 環境を用いて，代表的な協調行動の一つであるプレイヤ間のパスによるボールの受け渡しについて，パスを送るエージェントがパスを高い確率で成功させるためにとるべき行動の学習を強化学習を用いて行い，その有効性を確かめた．類似の研究としては Stone ら [1][2] がある．この研究は，味方チームでボールをキープし続けるというタスクを，ボールを持ったエージェントの行動を強化学習を用いて学習させ，キープする事の出来る時間を長くしていく，というものであった．本研究のタスクはこの研究に比べて単純なものであるが，学習エージェントの取りうる行動はこの研究に比べて多くの種類の行動を取る事が出来るように設定した．

2. 実験環境

実験環境として用いた RoboCup Soccer Simulator は，サーバとサッカークライアントが UDP/IP を用いて通信するモデルとなっている．1 つのクライアントが制御するのは，1 つのエージェントのみで，1 つのクライアントが複数のエージェントの情報を集中制御することは，原則として禁止されている．また，サーバを介さずに，クライアント同士が独自の方式を用いて通信することも禁止されている．また，不完全知覚，情報の不確実さ，実時間処理など現実の問題に近い課題を多く含む環境として設計されている．

3. 実験

3.1 実験設定

仮想フィールド上に 3 体のエージェントを配置し，実験を行った．各エージェントを Agent1,Agent2,Agent3 と呼ぶ．Agent1 が学習エージェントであり，Agent3 にボールを奪われないように Agent2 にボールをパスするための行動をパスの成功を報酬とした強化学習 [3] を用いて学習する．学習のアルゴリズムには Profit-Sharing を用いた．Agent1 が取り得る行動は以下の通りである．

- 前方にドリブルする．
- +90 度方向にドリブルする．
- -90 度方向にドリブルする．
- Agent2 にボールをパスする．
- 何もしない．

尚，ドリブルの方向はボールのある方向を 0 度とする．Agent2，Agent3 は定められたルールに従って動く組み込みのエージェントである．Agent2，Agent3 の行動パターンは以下のとおりである．

- Agent2:ボールが自分の 10m 以内に入ったとき，ボールを追いかける．それ以外のときはその場に静止する．
- Agent3:常にボールを追いかける．

実験の手順は以下の通りである．

1. エージェント，ボールの位置を初期座標へ初期化する．
2. 試行を開始する．
3. Agent1 から Agent2 へのパスが成功したとき，または Agent3 にボールを奪われたときを終端状態とし，パスが成功した時にだけ Agent1 は報酬を受け取る．試行開始から 10 秒経過したときも終端状態とし，このとき Agent1 がボールを持っていた場合も失敗とする．
4. 1 に戻り試行を繰り返す．

Agent1 から Agent2 へのパスが成功したとは，Agent2 がボールの半径 1.5m 以内に近づいたときにボールの半径 1.5m 以内に Agent3 がいない状態であったとき，する．Agent3 にボールを奪われたとは，Agent2 がボールの半径 1.5m 以内に近づいたときにすでに Agent3 がボールの半径 1.5m 以内にいた状態であったとき，とする．また，Agent1 がボールを持っている状態で，Agent3 が Agent1 の半径 1.0m 以内に近づいたときもボールを奪われた，とする．成功，失敗の判断は，Agent1 の観測により行うこととする．なお，本実験では簡単化のため本来 RoboCup Soccer Simulator に備わっている，エージェントの視野角の制限，スタミナの影響を無視するよう設定を変更して実験を行った．

†京都工芸繊維大学大学院 工芸科学研究科 Graduate School of Engineering,Kyoto Institute of Technology.

すべての $s \in S, a \in A$ に対して：
　$P(s,a) = C$(C は任意の小さな正の定数)
各エピソードに対して繰り返し:
　状態 s を初期化
　エピソードの各ステップに対して繰り返し:
　　P から導かれる重み付きルーレットを用いて，s での行動 a を選択する
　　行動 a を取り，報酬 r と次状態 s' を観測する
　　$s \leftarrow s'$
　s が終端状態ならば繰り返しを終了
エピソードに含まれる全ての状態行動対に対して:
$P(s_t, a_t) \leftarrow P(s_t, a_t) + f(t, r_T, T)$

図 1: Profit Sharing のアルゴリズム

3.2 Agent1 の学習アルゴリズム

強化学習に望まれる性能には，結果としてなるべく大きい報酬を得ようという最適性と，学習途中においてもなるべく報酬を得続けようという効率性の二つの側面がある．本研究で使用した Profit Sharing とは，後者を重視した学習手法の一つであり，報酬を得たときに，それまでに使用されたルール系列を一括的に強化する手法である．Profit Sharing のアルゴリズムを図 1 に示す．

図 1 において，$P(s,a)$ は状態 s における行動 a の優先度，f は信用割当関数と呼ばれる関数であり，Profit Sharing ではエピソード終了時にエピソードに含まれる各状態行動対 s_t, a_t に対する優先度 $P(s_t, a_t)$ をこれを用いて一括して強化する．

信用割当関数として，宮崎ら [4] によって考案された等比減少関数

$$f(t, r_T, T) = \gamma^{T-t-1} r_{T-1} \quad (0 \leq \gamma \leq 1) \qquad (1)$$

がよく用いられ，本論文でもこれを用いる．式 1 の γ は割引率と呼ぶ．

Profit Sharing においては，行動は優先度 P に比例した確率分布に従って選択する．すなわち，状態 s で行動 a が選択される確率は，

$$P_r(a_t = a | s_t = s) = \frac{P(s,a)}{\sum_{a' \in A(s)} P(s,a')} \qquad (2)$$

である．

今回行った実験では，報酬 $r = 1.0$，割引率 $\gamma = 0.9$，行動優先度 P の初期値 $C = 0.1$ と設定した．また，行動決定は単位時間毎ではなく，状態観測が変化する毎に行う，とした．

3.3 実験 1

以下の様に条件を設定し，実験を行った．

- エージェントおよびボールの初期座標は，Agent1 が (-20,0)，Agent2 が (5,0)，Agent3 が (-5,0)，ボールが (-19.5,0) と設定する (図 2(a))．
- 状態として使うのは，Agent1 から Agent2 までの距離，Agent1 から Agent3 までの距離，Agent1 の体の正面を基準とした Agent2 との角度，Agent3 との角度の 4 次元とする (図 3(a))．

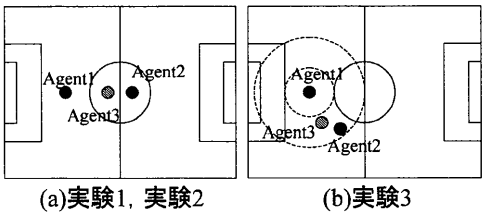

図 2: エージェントの初期座標

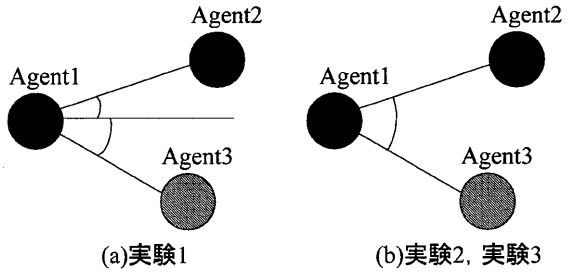

図 3: 状態表現

- エージェント間の距離に関しては，15m 未満までは 1m 毎に閾値処理により離散化，40m 未満までは 5m 毎に離散化，40m 以上は 1 つの状態として扱う，とする．
- Agent1 の体の正面から他のエージェントとの間の角度に関しては，0 度から 22.5 度毎に離散化する．

実験 1 の結果を図 4 に示す．学習の結果は，最初の 100 試行目までは 10 試行毎に，100 試行目から 1000 試行目までは 100 試行毎に学習を止め，その時点での行動優先度に従って 100 試行のパス実験を行い，パスの成功数をプロットしたものである．ランダムとは，全ての行動が等確率で選択される場合であり，その値については 100 試行を 10 回行い，成功数の平均を取ったものである．

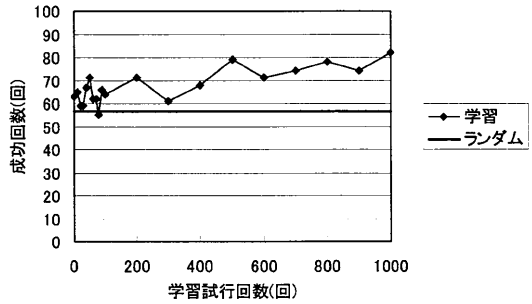

図 4: パスの成功回数の推移

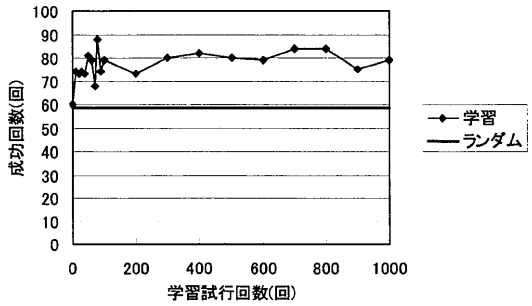

図 5: パスの成功回数の推移

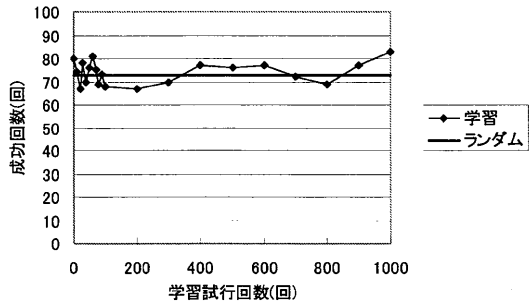

図 6: パスの成功回数の推移

3.4 実験 2

実験 1 から状態表現を以下の様に変更し，実験を行った．

- 状態として使うのは，Agent2 までの距離，Agent3 までの距離，Agent1- Agent2，Agent1-Agent3 の 2 線分間の角度の 3 次元とする (図 3(b))．
- 角度の離散化は 10 度毎に行う．

実験 2 の結果を図 5 に示す．

3.5 実験 3

実験 2 と同じ状態表現で，Agent2，Agent3 の初期座標を各試行毎に Agent1 の半径 5m 以上 20m 以内のランダムな位置に変更し (図 2(b))，同様の実験を行った．実験 3 の結果を図 6 に示す．

3.6 考察

実験 1 と実験 2 を比較すると，実験 2 の方が学習が早く進んでいる事がわかる．これは状態表現を変更する事により状態行動対の数を削減した結果であると考えられる．また，実験 3 ではより多くの状態が起こりうるようにするため Agent2，Agent3 の初期座標を試行毎にランダムに変更して実験を行ったが，この場合でも学習が多少は進んでいるとも取れるが，改良の必要があると考える．

4. 結び

本研究の手法を用いることにより，パスに至るまでの行動を学習によって獲得できることを確認できた．今後は適応的な状態空間の分割や，複数のエージェントが同時に学習する環境への対応等について実験を行いたいと考える．

参考文献

[1] Peter Stone, Richard S. Sutton: "Scaling Reinforcement Learning toward RoboCup Soccer", In Proceedings of the Eighteenth International Conference on Machine Learning, pp. 537-544 (2001)

[2] Gregory Kuhlmann, Peter Stone: "Progress in Learning 3 vs. 2 Keepaway", RoboCup-2003 (2003)

[3] Richard S. Sutton,Andrew G.Barto 著, 三上貞芳, 皆川雅章 訳: "強化学習", 森本出版, 第 1 版 (2000)

[4] 宮崎和光，山村雅幸，小林重信: "強化学習における報酬割当ての理論的考察"，人工知能学会誌,Vol.9,No.4,pp.580-587 (1994)