社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-A901-1 (5/9)

RoboCup サッカーの Keepaway サブタスクにおけるパスの受け手の強化学習

Reinforcement Learning of Receivers for RoboCup-Soccer Keepaway

坂本 聖也　尾関 基行　岡 夏樹
Seiya SAKAMOTO　Motoyuki OZEKI　Natsuki OKA
京都工芸繊維大学
Kyoto Institute of Technology
m9622015@edu.kit.ac.jp

Abstract

In this paper we discuss reinforcement learning of "the keepers" in the keepaway subtask of RoboCup soccer. The keepers try to keep control of the ball as long as possible despite the efforts of the takers in a 20m by 20m region. While the most of previous studies focus on a keeper who makes a pass, we focus on receivers. The receivers learn their policy by reinforcement learning using the degrees of congestion as the state variables. In the experiments, we first check the proper positions where the degrees of congestions are calculated. We then compare our method with a hand-coded policies under two taker's policy—smart and normal, and show that our keepers can keep a ball longer than hand-coded ones under both taker's policies.

1　緒言

エージェントとは、ある環境を知覚し、自分で判断して行動するものである。しかし、単一のエージェントの能力には限界があり、大規模かつ複雑な問題を扱うのは困難である。そこで、単一のエージェントでは対処困難な問題に、複数のエージェントで取り組むことによってその問題を解決しようとするマルチエージェントシステム（Multi-Agent System：以下、MAS）が注目されている。MAS 全体の振る舞いは、エージェント同士の相互作用によって決定される。よって MAS では、各エージェントの能力以上にエージェント同士の協調行動が重要となる。

MAS のテストベッドとして RoboCup Soccer Simulator[1]（以下、RCSS）というサッカーを例題としたものがある。サッカーのチームは 11 人の選手で構成された MAS であると考えることができる。つまり、RCSS でサッカーエージェントを設計する際も、エージェント同士の協調行動を考慮することが重要である。しかし、サッカーのような多数のエージェントが存在する環境においてすべての行動を設計するのは困難なため、学習によって適切な行動を獲得するアプローチが期待されている。

RCSS を用いた機械学習に関する研究は多いが、ボールを持ったエージェント（パスの出し手）に着目したものが多く、ボールを持っていないエージェント（パスの受け手）の機械学習に関する研究は手つかずである。MAS であることを考えると、パスの受け手の行動や出し手と受け手との協調行動も重要である。そこで本研究では、パスの受け手の学習に着目し、混雑度を状態変数とした強化学習の手法を提案する。

以下、第 2 章では Keepaway タスクの問題設定と従来研究について述べる。第 3 章ではパスの受け手の学習方法を説明する。第 4 章では実験の設定と手順を説明する。第 5 章では実験の結果と考察を述べる。最後に第 6 章で本研究をまとめ、今後の課題を述べる。

2 Keepaway タスク

2.1 問題設定

本研究では RCSS の中で強化学習などの繰り返し試行に適した Keepaway[2] というサブタスクを用いる。Keepaway は、20m × 20m の限られた領域内で 3 人の keepers と呼ばれるチームが 2 人の takers と呼ばれる相手チームにボールを取られないようにパスを回し続ける連続タスクである。keepers チーム、takers チームの各エージェントをそれぞれ、keeper、taker と呼ぶ。keeper はさらにパスの出し手とパスの受け手に分けることができる。takers がボールを奪うか、ボールが領域の外に出るとエピソードが終了し、すぐにまた次のエピソードが始まる。各エピソード開始時のエージェントは、Figure 1 のように配置される。2 人の takers は左下の隅に、keepers は残り 3 つの隅に初期配置され、ボールは左上の keeper が持った状態で開始される。

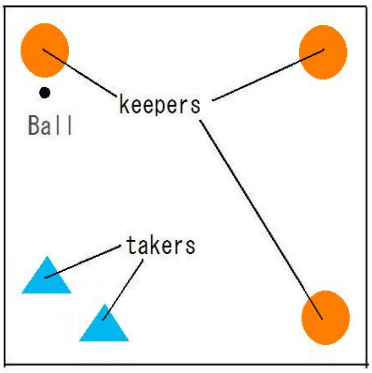

Figure 1: A starting configuration for a Keepaway episode.

2.2 従来研究

Stone ら [3][4] や荒井ら [5] の行った研究では、パスの出し手の行動選択肢はボールを保持し続けるか味方の keeper へパスするかの 2 つであり、伊佐野ら [6] の研究ではこれにドリブルを追加導入した。しかし、いずれの研究も学習するのはパスの出し手の行動のみであった。

報酬については、Stone ら [3][4] はステップごとにプラスの報酬を、荒井ら [5] はエピソード終了時に終了する原因となったエージェントにマイナスの報酬を与えた。荒井ら [5] の報酬設計の方が優れていたため、伊佐野ら [6] は後者の報酬設計を用いた。

これらの研究では、状態変数として Figure 2 のように距離変数 11 個と角度変数 2 個（合計 13 個）を用いて学習を行っている。Figure 2 ではボールを持っている keeper を k1、ボールを持っていない keeper を k2、k3 とし、k3 に近い方の taker を t1、k2 に近い方の taker を t2、さらに領域の中心を center point としている。

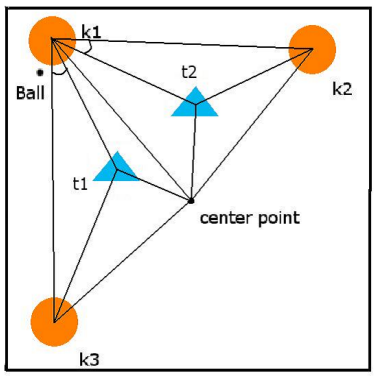

Figure 2: The state variables used in several previous studies.

3 学習方法

3.1 状態変数

本研究では強化学習で使用する状態変数を前述の従来研究から変更する。従来研究では keepers、takers、領域の中央との距離や角度を状態変数としていたが、本研究では領域の四隅と中央の混雑度を状態変数とする。混雑度とは、四隅と中央の座標から自分以外のエージェントまでの距離の逆数を合計したものである。

また、混雑度は連続値をとるため、1 次元タイルコーディングを行った。5 つの各変数ごとに 32 本のタイリングを用いて、各タイリングの分割数を 10 とした。

本研究での状態変数は Figure 3 のようになる。

3.2 行動

keepers はまずボールを探し、3 人の内ボールに最も近いエージェントがボールを取りに行く。そし

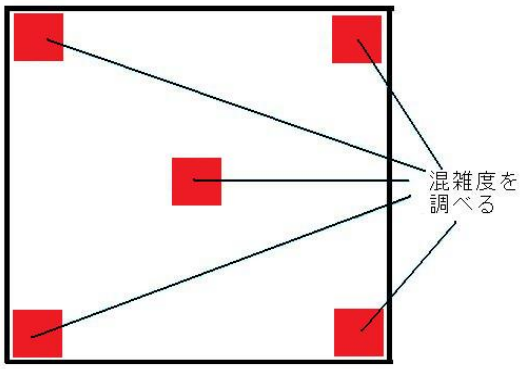

Figure 3: The state variables used in this article.

て、他の 2 人のエージェントがパスの受け手となり、状態変数である 5 点（四隅と中央）の混雑度から 1 点を決め、移動する。ただし、keepers と takers は、ボールが近くに来ると取りに行くように設計されており、移動の途中でもパスなどによってボールが近づくと、ボールを取りに行く。

3.3 報酬

報酬は、パスの受け取りに成功したとき（パスの受け手からパスの出し手に変わったとき）に与えることにする。本来はパスの受け手への報酬は目標位置に移動できた場合に与えたいが目標位置に移動できたとしてもトラップミスなどによりボールをうまく受け取れない場合もあるためにこのように決定した。

また、ボールを受け取らなかったもう 1 人のエージェントにも、直接ではないがボールキープ行動に関わったと考え、同じ大きさの報酬を与える。

3.4 Q-learning

本研究ではパスの受け手の学習に Q-learning を用いる。Q-learning は有限マルコフ決定過程において、全ての状態が十分にサンプリングできるようなエピソードを無限回試行した場合、最適解に収束することが証明されている。

Q-learning では行動価値である Q 値を式 (1) のように更新する。

$$Q(s,a) \leftarrow Q(s,a)+\alpha\left[r+\gamma \max_{a' \in A(s')} Q(s',a') - Q(s,a)\right] \quad (1)$$

式 (1) において、$Q(s,a)$ は状態 s において行動 a をとる時の価値で、$Q(s',a')$ は s の次状態 s' において行動 a' をとる時の価値である。また、r は報酬、$A(s')$ は状態 s' で実行可能な行動全体の集合、$\alpha(0 \leq \alpha \leq 1)$ は学習率、$\gamma(0 \leq \gamma \leq 1)$ は割引率を表している。ここで Q 値の更新は、エージェントが行動を実行し、次の状態に移るたびに行われる。

次に Q-learning のアルゴリズムを示す。

1. Q 値を初期化する。
2. エピソードの繰り返し数を設定する。
3. 状態 s を初期化する。
4. 状態 s で実行する行動 a を Q 値から決定する。
5. 行動 a を実行し、報酬 r があれば獲得、その後、次状態 s' へ遷移する。
6. Q 値を式 (1) に従って更新する。
7. エピソードが終了しなければ 4. へ。
8. エピソードの繰り返し数に達すれば終了。そうでなければ 3. へ。

アルゴリズムの 4. の行動選択法には様々な方法があるが、本研究では ϵ -greedy 法を用いる。これは選択確率 $\epsilon(0 \leq \epsilon \leq 1)$ を決め、1-ϵ の確率で Q 値が最大の行動を、ϵ の確率でランダムな行動を選択する方法である。

4 実験

4.1 実験設定

本実験では、パスの出し手及び takers は学習しないこととする。パスの出し手は、takers が自分から 5m までの範囲にいれば味方の keeper へのパスを選択し、5m より離れていればホールドボールを選択する。味方へのパスを選択したときは、4.0 × (味方の keeper と近い方の taker との距離)+(味方の keeper と近い方の taker との角度) を二人の味方の keeper それぞれへのパスの成功しやすさとし、パスが成功しやすい方へパスをする。takers の戦略には A と B の 2 種類を用いる。戦略 A では 2 人ともボールを追いかける行動を実行し続ける。戦略 B ではボールに近い方がボールを追いかけ、もう 1 人はパスの受け

手 2 人のうちパスの出し手から離れている方をマークする。戦略 B では 2 人が役割分担するため戦略 A よりも高度なものである。

実験で使用するパラメータを Table 1 に示す。

Table 1: Parameter values used in the experiments.

報酬 r	1
学習率 α	0.1
割引率 γ	0.95
選択確率 ε	0.1

4.2 実験手順

本稿では以下の 2 つの実験を行った。

実験 1

本研究では領域の四隅と中央の混雑度を状態変数にするとした。しかし、Figure 4 のように、中央の座標 (center) は (10,10) と 1 つに決められるが、四隅は 1 つには決められない。keepers が領域をいっぱいに使えば、takers にボールを取られにくくなるが、ボールが領域の外に出る可能性が高くなる。逆に内側に行き過ぎると、ボールは領域の外に出にくくなるが、takers にボールを取られる可能性が高くなる。そこで、まず最も良い点を実験により探すことにする。領域の 4 つの各頂点をはじめとして、x 座標、y 座標をともに 1m ずつ内側にずらしていき、5 通り調べる。1 つ目は、(0,0), (0,20), (20,0), (20,20)（以下、(0,20) と表記する）、2 つ目は、(1,1), (1,19), (19,1), (19,19)（以下、(1,19) と表記する）、3 つ目は、(2,2), (2,18), (18,2), (18,18)（以下、(2,18) と表記する）、4 つ目は、(3,3), (3,17), (17,3), (17,17)（以下、(3,17) と表記する）、5 つ目は、(4,4), (4,16), (16,4), (16,16)（以下、(4,16) と表記する）である。それぞれについてパスの受け手を takers の戦略 A と B の場合に分け、Q-learning で学習させる。

実験 2

実験 1 で最も良い結果が得られた座標において、takers の戦略 A と B の学習結果の比較、及び学習後のエピソード平均持続時間の手コードされたものとの比較を行う。手コードは Stone ら [3][4] が定義したマクロ行動の中の GetOpen という行動（パスの受け手は takers から離れて、味方からのパスを受け取れる位置に移動する行動）を実行する。

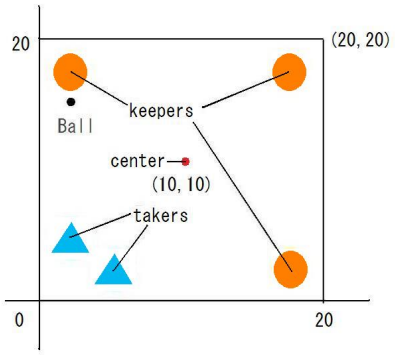

Figure 4: The coordinate system of the Keepaway region.

5 結果・考察

5.1 実験 1

Figure 5 に戦略 A の 5 通りの四隅それぞれについてのパスの受け手の学習結果を示す。Figure 5 のエピソード持続時間とは、30 エピソード毎の平均持続時間である。すべてのデータをプロットすると分かりにくくなるので、2 時間ごとのデータのみプロットした。

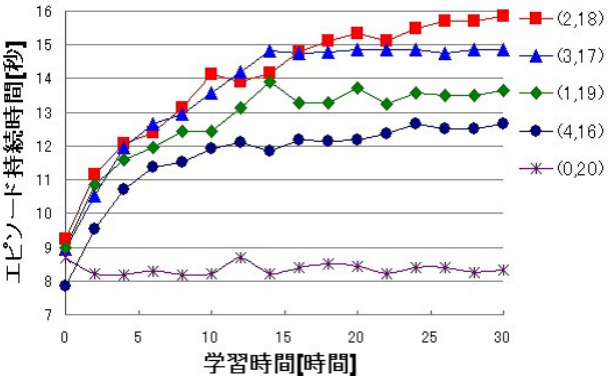

Figure 5: Learning curves for taker's policy A.

まず、Figure 5 をみると、(0,20) については全く学習ができていない。この理由としては、keeper は (0,20) という領域の境界線上におり、ボールが領域の外へ非常に出やすくなっているため、時間が伸びなかったと考えられる。(0,20) よりも内側のものについては学習できていることが確認できる。

30 時間学習させた後の 30 エピソード毎のエピソード平均持続時間 30 個を用いて、1 番良い結果と思われる (2,18) と 2 番目に良い結果と思われる (3,17) の間で t 検定した結果、(2,18) の方が (3,17) よりも優れていることが分かった（t(29)=61.60, p<.05)。

なお、(2,18) 以降は (3,17)、(4,16) と内側に行くにつれてエピソード持続時間は短くなっていく傾向がみられるので、これ以上内側を調べてもさらに時間は落ちることが予想される。

次に、Figure 6 に戦略 B の 5 通りの四隅それぞれについてのパスの受け手の学習結果を示す。Figure 6 のエピソード持続時間とは、30 エピソード毎の平均持続時間である。Figure 5 と同じ縮尺にすると分かりにくくなるので、縦軸の縮尺を変更した。すべてのデータをプロットすると分かりにくくなるので、2 時間ごとのデータのみプロットした。

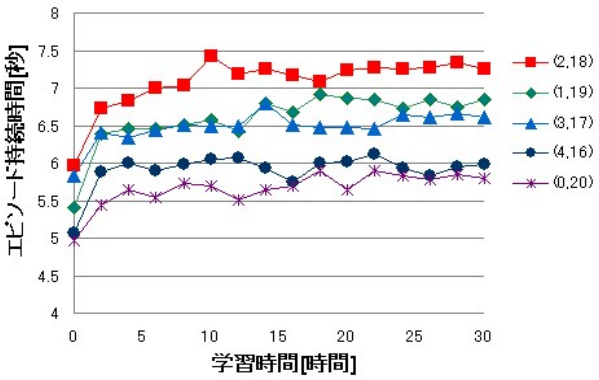

Figure 6: Learning curves for taker's policy B.

30 時間学習させた後の 30 エピソード毎のエピソード平均持続時間 30 個を用いて、1 番良い結果と思われる (2,18) と 2 番目に良い結果と思われる (1,19) の間で t 検定した結果、(2,18) の方が (1,19) よりも優れていることが分かった（t(29)=40.81, p<.05)。

5.2 実験 2

実験 1 で takers の戦略 A、B ともに最良の結果となった (2,18) において、takers の戦略 A、B の学習結果の比較を行った。Figure 7 に takers の戦略 A、B の学習結果を示す。すべてのデータをプロットすると分かりにくくなるので、2 時間ごとのデータのみプロットした。

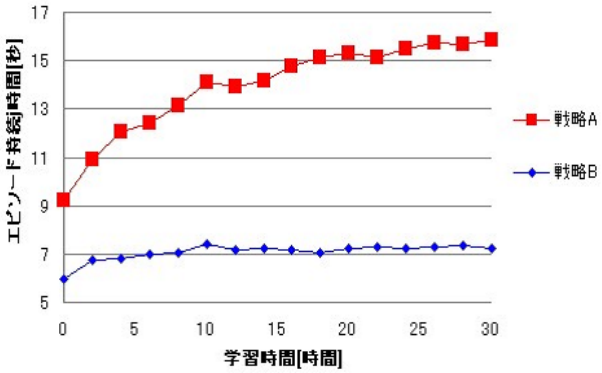

Figure 7: Learning curve comparison for taker's policy A and B.

戦略が高度な戦略 B は学習後も 1 秒程度しか時間は伸びておらず、戦略 A と比較すると学習の成果はあまりない。エピソード持続時間も半分以下となり、takers の戦略が結果に大きな影響を与えることが分かった。

最後に、手コードと 30 時間学習させた後の (2,18) のエピソード平均持続時間を示す。

Table 2: Comparison of average durations for hand-coded and learned policies.

	手コード	(2,18)
戦略 A	14.9	15.8
戦略 B	7.1	7.3

30 時間学習させた後の 30 エピソード毎のエピソード平均持続時間 30 個を用いて手コードと (2,18) の間で t 検定を行った結果、戦略 A、B 共に有意差があることが確認できた（A：t(29)=-32.92, p<.05、B：t(29)=-21.08, p<.05)。したがって、戦略 A、B 両方とも学習させた方が手コードよりも優れているといえる。

6 結言

本研究では RCSS の Keepaway タスクを使用し、ボールを持っていない味方エージェントであるパスの受け手に注目し、パスの受け手の学習を行った。そして、takers の戦略が A、B どちらの場合も、(2,18) で最も良い結果が出て、手コードよりも優れていることが分かった。

本研究では (2,18) で最も良い結果が出たが、全ての点を調べたわけではないので、もっと細かく調べると、より良い結果が得られる点があるかもしれない。また、本研究で用いた状態変数や、行動の種類を変えたりすると、もっと良い結果が出るかもしれない。強化学習の方法や報酬設計についても同じことが言えるだろう。

また、本研究により、takers の戦略が時間や学習結果に大きな影響を与えることが分かった。

MAS で非常に重要なことは協調行動であるので、パスの受け手だけでなく、今後はパスの出し手と受け手の両方を学習させたり、takers も学習させることが必要である。

今回用いた Keepaway サブタスクは RoboCup サッカーの一部にすぎないので、今後は実際の試合である 11 人対 11 人を想定した研究をしなければいけない。

謝辞

本研究を行うにあたり、テーマ設定について助言をいただいた野田五十樹氏に心より感謝いたします。

参考文献

[1] RoboCup Soccer Simulator
http://sserver.wiki.sourceforge.net/

[2] Learning to Play Keepaway
http://www.cs.utexas.edu/~AustinVilla/sim/keepaway/

[3] Stone, P. and Sutton, R. S.
"Reinforcement Learning toward RoboCup Soccer"
in Proceedings of 18th International Conference on Machine Learning, pp.537-544, (2001)

[4] Peter Stone, Richard S. Sutton, and Gregory Kuhlmann.
"Reinforcement Learning for RoboCup-Soccer Keepaway"
Adaptive Behavior, Vol.13, No.3, pp.165-188, (2005)

[5] 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"
人工知能学会論文誌, pp.537-546, (2006)

[6] 伊佐野勝人, 片上大輔, 新田克己
"ロボカップサッカーシミュレーションの Keepaway における協調行動の学習"
第 22 回人工知能学会 全国大会, 2i3-03, (2008)

[7] 高玉圭樹
"マルチエージェント学習"
コロナ社, (2003)

[8] 秋山英久
"ロボカップサッカーシミュレーション 2D リーグ必勝ガイド"
秀和システム, (2006)