F-031

RoboCupSoccer における Q 学習を用いた対人守備の学習
An Algorithm to Train Defense Using Q-learning for RoboCup Soccer

水島 諒† 穴田 一‡
Ryo Mizushima Hajime Anada

1. はじめに

近年，「ゲーム AI」の開発が盛んに行われている．例えば，チェスや将棋，囲碁といったゲームが挙げられる．そして，これらのゲームにおいては AI が人間のチャンピオンに勝利するといった事も起きている．また，RoboCup と呼ばれる，自律型ロボットによるサッカーの世界大会が毎年行われている[1]．RoboCup とは，西暦 2050 年迄にサッカーの世界チャンピオンチームに勝てる，自律型ロボットのチームを作ることを目標とした大会である．この RoboCup には 5 つのリーグがあり，リーグごとに異なる特徴がある．本研究では 5 つのリーグのうち，各選手がそれぞれ思考し，人間のような戦術的なサッカーが行われている 2D リーグを扱う．2D リーグは，移動可能範囲が広いことや，リアルタイムに計算し判断を下す必要があること，11 人同士の対戦であることから，前述のチェスや将棋，囲碁といったゲームより難しいと考えられている．そして，秋山は RoboCup の 2D リーグで使用可能な agent2d (Ver 3.1.1) というチームモデルを公開している[2]．このチームモデルでは，全てのエージェントがボールの位置のみを使用し，移動先を決定している．しかし，このモデルはどのような戦況においても同じポジショニングを目指すという問題がある．そこで，本研究では戦況に応じた守備を行うため，サッカーの 2 対 2 の練習メニューを取り入れ，Q 学習を用いてエージェントにボールを奪う守備を学習させる方法を提案する．

2. 既存研究

2.1 秋山のチームモデル（agent2d）

秋山は，エージェントの移動先の決定を行うためにフォーメーションシステムを提案した[3]．フォーメーションシステムでは，全てのエージェントがボールの位置のみを用いて移動先を決定している．このシステムでは，事前にフィールド上の複数の位置に，それらの位置にボールがあった場合の 11 人のエージェントの最適位置を，それぞれ設定している．

2.2 1 対 1 の状況における守備の学習

著者は，agent2d の問題点である「どのような戦況においても同じポジショニングを目指す」，「ボールを奪う動きが少ない」を解決するため，Q 学習を用いて 1 対 1 の状況におけるボールを奪う守備を学習している[4]．これにより，練習メニューにおける学習はうまくいったが，試合に適用した際，弱くなるという結果になった．

† 東京都市大学大学院 工学研究科
‡ 東京都市大学 知識工学部

3. 提案するチームモデル

秋山が開発したチームモデルである agent2d (Ver 3.1.1)は，「どのような戦況においても同じポジショニングを目指す」，「ボールを奪う動きが少ない」という問題があった．その問題を解決するために，著者は 1 対 1 の練習メニューを導入した[4]．しかし，試合に適用した際に弱くなってしまうという結果になった．実際のサッカーでは周りと協調して守備を行う．そこで，練習メニューを拡張し，ボールを奪いに行く選手のフォローを行う選手の守備方法を学習する．学習方法は，Q 学習を用いて 2 対 2 の練習を行い，ボールを奪う守備を学習させる．守備の練習メニューとしては，実際のサッカーで使用される練習メニューを用いる．

3.1 練習メニュー

実際のサッカーでは，2 対 2 の状況の練習メニューとして，図 2.1 のようなものがある．練習メニューのフィールドは，長辺の長さが 20 m，短辺の長さが 15 m である．● は攻撃側のエージェントであり，左の赤色のラインの外にドリブルやパスを行い，ボールを出すよう攻める．● は守備側のエージェントであり，左のライン以外の青色のライン外にボールを出すように守るという練習方法である．

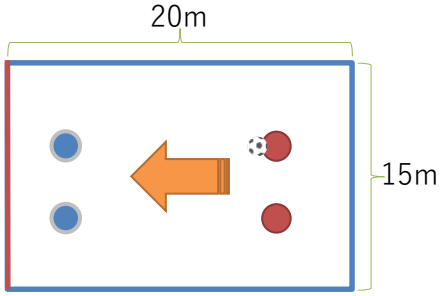

図 2.1 2 対 2 の練習メニュー

Q 学習で使用するため，練習メニューで用いるフィールドの座標系を図 2.2 のようにした．赤色ライン中央を原点とし，長辺方向を x 軸，短辺方向を y 軸とした直交座標系で表す．

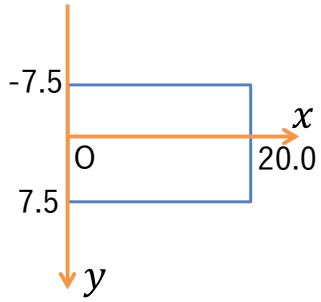

図 2.2 2 対 2 の練習メニューで用いるフィールドの座標系

3.2 Q 学習の適用方法

3.2.1 Q 学習の概要

本研究で用いる Q 学習の基本的な枠組みを述べる．Q 学習では，エージェントは現在の環境の状態 S_t を観測し，実行すべき行動 a を選択する．行動 a による環境の変化に応じた報酬 r を受取り，環境の状態は S_{t+1} に変化する．その時，次式を用いて状態 S_t における行動 a の行動価値 $Q(S_t, a)$の更新を行う．

$$Q(S_t, a) \leftarrow Q(S_t, a) + \alpha \left[r + \gamma \max_p Q(S_{t+1}, p) - Q(S_t, a) \right] \quad (1)$$

ここで，α は学習率，r は報酬の量，γ は割引率を表し，$\max_p Q(S_{t+1}, p)$ は状態 S_{t+1} における可能な行動の中の最大の行動価値を表す．エージェントは，観測と行動選択を繰り返すことにより，守備に有効な行動に対する行動価値 $Q(S_t, a)$を更新していく．

Q 学習を用いるために，エージェントの報酬 r，エージェントの観測できる状態 S_t と選択できる行動 a は後述で詳しく述べる．

3.2.2 報酬の設計

良い行動をした時には目的の達成度合に応じた報酬を与える．本実験は，図 2.1 の青色のライン外にボールを出し，赤色のライン外にボールを出されないことが目的である．そこで，この目的を達成するために次の 3 つの小目的を設定した．

I. ボールを奪う
II. コート外に出ない
III. ボールを後方に戻させる

この 3 つの小目的それぞれの達成具合を表す報酬 r を次式のように定義した．

$$r = \sum_{i=1}^{3} reward_i \quad (2)$$

$$\begin{cases} reward_1 = \begin{cases} 10000 & if\ Take\ the\ ball \\ 0 & otherwise \end{cases} \\ reward_2 = \begin{cases} -100 & if\ Out\ of\ the\ field \\ 0 & otherwise \end{cases} \\ reward_3 = \max(x_b - prebx_b, 0.0) \times 10.0 \end{cases}$$

ここで，$reward_i$ 小目的 i に応じた報酬を表す．x_b はボールの x 座標，$prebx_b$ は 1 秒前のボールの x 座標を表す．図 2.3 に報酬のための変数設定を示す．

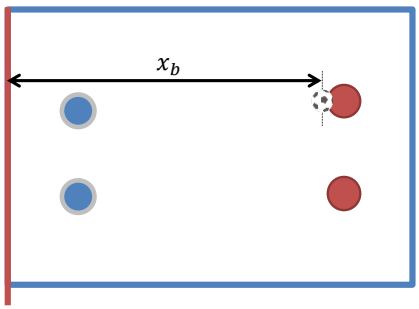

図 2.3　報酬で用いる変数設定

3.2.3 状態の定義

守備を行う際，自分や味方，敵の位置，ラインからの距離などを考慮して行動を選択しなければならない．そこで，守備行動を学習するための状態を次のように定義する．

- ボール保持者が自分から見て 8 方向のどの方向か
- 保持者ではない敵が 8 方向のどの方向か
- 保持者でない敵が自分からP_1 m 以内にいるか
- 保持者でない敵が自分からP_2 m 以内にいるか
- ボール保持者が自分からP_1 m 以内にいるか
- 味方がボール保持者からP_1 m 以内にいるか
- 保持者でない敵がボール保持者からP_1 m 以内にいるか
- 図 2.1 の青色のラインからボールまでの距離がP_1 m 以内か

以上のように状態を定義すると，エージェントが識別する状態は 4096 種類となる．P_1，P_2はパラメータであり，P_2 はP_1の 2 倍の値である．

3.2.4 行動の選択

エージェントが可能な行動は 8 方向に歩く，走る，又はタックル，動かない，計 18 種類とする．学習中の行動の選択は，よりよい行動を探すために，ε の確率でランダムに行動を選択し，$(1-\varepsilon)$ の確率でそれまでに獲得した最適な行動を選択するように設定した．

4. モデルの評価

本研究では，学習率 α は 0.1，割引率 γ は 0.9，行動の選択確率 ε は 0.3 とした．攻撃側のエージェントの 2 人の初期配置は，(18.0, 4.0)，(18.0, -4.0) の位置を中心とし，半径 3 m 以内の範囲内にランダムにそれぞれ配置する．そして，守備側のエージェントの 2 人の初期配置は，(2.0, 4.0)，(2.0, -4.0)の位置を中心とし，半径 3 m 以内の範囲内にランダムにそれぞれ配置する．ボールは攻撃側エージェントのどちらかに渡す．また，攻撃側のエージェントには，既存研究のチームモデルである agent2d (Ver 3.1.1)のエージェントを使用した．守備側のエージェントは，ボール保持者から一番近い時，事前に 1 対 1 の練習メニューで学習した守備を行う[4]．それ以外の時は，提案した学習を行う．

学習方法は，30 秒間練習メニューを用いた練習を行い，それ迄にコート外に出すことが出来なかった際は，初期配置に戻し，練習を行う．また，30 秒以内にコート外にボールが出た場合も同様に初期配置に戻し，練習を行う．これを行動価値 $Q(S_t, a)$ が収束するまで繰り返す．

発表時にモデルと詳細な計算結果と考察を述べる．

参考文献

[1] Hiroaki Kitano, Minoru Asada, Yasuo Kuniyoshi, Itsuki Noda, Eiichi Osawa and Hitoshi Matsubara, “RoboCup: A Challenge Problem for AI”, AI Magazine , Vol.18, No.1, pp.73-85, (1997).
[2] 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).
[3] Hidehisa Akiyama, Hiroki Shimora, and Itsuki Noda: HELIOS2009 Team Description. In Robocup 2009, (2009).
[4] 水島 諒, 穴田 一, “RoboCup における Q 学習を用いた守備の強化”, 人工知能学会研究会資料, SIG-SAI, Vol.26, No.1, pp.1-6 , (2016)