7P-05

RoboCup サッカーシミュレーション 2D リーグの守備における評価関数の設計と学習

山中琢矢† 五十嵐治一‡
芝浦工業大学†

1. はじめに

RoboCup サッカーシミュレーション 2D リーグ[1]はコンピュータの仮想フィールド上でサッカーを行う競技であり，サンプルプログラムとして agent2d が提供されている[2].

agent2d の守備方法は，ボールの座標に依存する守備パターンに基づいて，守備プレイヤの位置を決定しており，相手プレイヤの位置を考慮していない．そのため，相手チームのパスやドリブルに対応できないことがある．そこで，入倉らは敵プレイヤのボール保持者と，敵レシーバを予測し，その 2 種のプレイヤと味方プレイヤとの距離の総和が最も小さくなるように守備を割り当てプレスやマークを行った[3].

また，agent2d はボール保持者の行動決定時に行動連鎖の探索木を用いた「チェーンアクション」を提案している．一方で，ボール非保持者の移動先は，前述のとおりボールの位置とホームポジションを基に計算されている．そこで，大内らはチェーンアクションを利用したボール非保持者の移動先決定方式を提案し，そのための局面評価関数の提案と学習を行った [4].

本研究では，守備の割り当てには入倉の方法を利用し，その後のボール非保持者の移動行動については大内の決定法を利用する．この際，チェーンアクションには新たに設計，学習した評価関数を用いることで守備力の向上を図った.

2. 本研究で提案する守備方式

2.1 対象とする守備の範囲と行動

agent2d-3.1.1 ではボール非保持者の行動はタックル，インターセプト，移動の 3 種類であるが，本研究では移動のみを対象とする.

Design and learning evaluation functions for defense in RoboCup Soccer Simulation 2D League
† Takuya Yamanaka・Shibaura Institute of Technology
‡ Harukazu Igarashi・Shibaura Institute of Technology

座標軸の原点をフィールド中央に取り，敵ゴール中心への方向をx方向($-52.5 < x < 52.5$)，それに直角な右方向をy方向とする．本研究では，ボールのx座標が$x < 10$の場合に，守備方式を適用する．これは，敵陣深くでのマークやプレスなどの守備は対象外と考えたためである.

2.2 守備の手順

以下の手順で守備行動を行う.
1．守備の割り当てを行う.
2．守備プレイヤの移動先候補を生成する.
3．チェーンアクションにより，移動先を決める.

3. 評価関数の設計

本研究では，守備プレイヤの移動先候補の決定のために次の評価関数$E(s;\omega)$を提案する.

$$E(s;\omega) = \sum_{i=1}^{3} \omega_i U_i(s) \quad (1)$$

sは評価対象となる局面，$U_i(s)$は評価項目，ω_iは$U_i(s)$の重みを表す．$U_i(s)$は守備プレイヤの「割り当てタイプ」によって変更する．割り当てタイプは以下の 3 種類である.

① 敵ボール保持者が割り当てられている.
② 敵ボール非保持者が割り当てられている.
③ 敵プレイヤが割り当てられていない.

①～③の場合の評価項目を表 1～表 3 に示す.

表 1 ①の場合の評価内容

評価項	評価内容
$U_1(s)$	自身と味方ゴールとの距離
$U_2(s)$	プレスのしやすさ
$U_3(s)$	割当て先の敵プレイヤとゴールの間にいるか

表 2 ②の場合の評価内容

評価項	評価内容
$U_1(s)$	自身と味方ゴールとの距離
$U_2(s)$	マークのしやすさ
$U_3(s)$	割当て先の敵プレイヤとゴールの間にいるか

表 3　③の場合の評価内容

評価項	評価内容
$U_1(s)$	自身と味方ゴールとの距離
$U_2(s)$	自身とホームポジションとの距離
$U_3(s)$	シュートコースの広さ

4. 評価関数の学習

4.1 学習則

学習には方策勾配法[5]を用いる．そこでは，エピソードを定義し，エピソード終了時点の状態やエピソード全体を評価し，報酬を与える．報酬の期待値を最大化するために，確率的勾配法を用いて評価関数のωを更新する．学習則は以下の(2),(3)で表される．学習中はBoltzman分布による方策(4)を用いる．

$$\Delta\omega = \varepsilon \cdot r \cdot \sum_{t=1}^{L} e_\omega(t) \tag{2}$$

$$e_\omega(t) \equiv \frac{\partial}{\partial\omega} \ln \pi(a_t|s_t;\omega) \tag{3}$$

$$\pi(a_t|s_t;\omega) = \frac{e^{E(s_t,a_t;\omega)/T}}{\sum_x e^{E(s_t,x;\omega)/T}} \tag{4}$$

ここで，s_tは時刻tにおける局面，a_tは時刻tで選択された行動，L はエピソード長，εは学習係数である．

4.2 報酬関数

守備時のエピソードに対する報酬rを表 4 に示すr_1~r_3の和として与えた．r_1は味方ゴールから遠いほど高く評価する項，r_2はエピソード長が短いほど高く評価する項，r_3はエピソードが終了した理由(敵に得点される等)を評価する項である．

表 4　報酬関数

報酬	評価内容
r_1	エピソード開始時と終了時の味方ゴールまでの距離の差に応じて 0~10.
r_2	エピソード開始時から終了するまでのサイクル数に応じて 0~10.
r_3	エピソードが終了した理由に依存して以下のように与える． ・敵に得点される．(-20) ・敵からボールを奪う．(+20) ・敵/味方が反則を行う．(+15/+10) ・敵/味方がボールを外に出す．(+15/+10) ・味方のゴールキックになる．(-10)

5. 実験

本研究ではセンターバック 2 人，サイドバック 2 人，ディフェンシブハーフ 1 人，オフェンシブハーフ 2 人の守備プレイヤに対して，学習を行った．重みの初期値は全て 1 である．学習前と学習後のチームそれぞれが agent2d と 500 試合行った結果を表 5 に示す．

表 5　agent2d との対戦結果

	勝率	勝-敗-分	得点	失点
agent2d	48.9%	197-206-97	2.32	2.33
学習前	0.4%	2-498-0	2.20	10.46
学習後	**62.3%**	248-150-102	2.19	**1.70**

また，r_1~r_3の内，どれが最も効果的であったかを確認するために，それぞれ単独で学習を行った．学習後に agent2d と 500 試合行った結果を表 6 に示す．

表 6　agent2d との対戦結果

報酬	勝率	勝-敗-分	得点	失点
r_1	50.5%	205-201-94	2.24	2.17
r_2	48.1%	205-221-74	2.37	2.42
r_3	**64.8%**	243-132-125	2.22	**1.80**

6. 結論

本研究では，マークとプレスの割り当て後に移動先を決定する際に，チェーンアクションを用いた．その際に，新たに設計，学習した評価関数を用いることで，agent2d の守備力の向上を図った．学習後，失点は 2.33 から 1.70 まで減少し，勝率を約 62%まで上昇できた．ただし，強化学習の報酬は守備のエピソードの終了時の状態に応じて与えるのが効果的である．

今後は，評価関数の見直しやヒューリスティクスを用いない関数近似を行うこと，強化学習で用いる報酬関数の改良などが課題である．

参考文献

[1] **ロボカップ日本委員会** Official Homepage http://www.robocup.or.jp/original/about.html(**参照日** 2020 **年** 12 **月** 26 **日**)

[2] RoboCup tools / soccer simulation wiki https://ja.osdn.net/projects/rctools/releases(**参照日** 2020 **年** 12 **月** 26 **日**)

[3] 入倉雅春,”RoboCup サッカーシミュレーションリーグ 2D における守備力の向上”, 芝浦工業大学修士論文(2018 年度)

[4] 大内斉,”ボール非保持者の移動先決定：局面評価関数の提案と学習”, 芝浦工業大学卒業論文(2015 年度)

[5] 五十嵐治一他,“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