ゲーム情報学 4－8
（2000．10．26）

RoboCupサッカーにおける局所的目標別行為－価値関数の設計

入沢 達矢　乾 伸雄　小谷 善行
irisawa@fairy.ei.tuat.ac.jp, [nobu, kotani] @cc.tuat.ac.jp
東京農工大学

概要

RoboCup サッカーにおいて困難な問題の 1 つに有効な行動を迅速に決定することがある。本論文では、行為－価値関数を用いることによってこの問題を解決している。しかし、サッカーのような複雑な環境において局面全体で有効な行為－価値関数を設計することは困難である。そこで本論文では、ボールの位置により局所的目標を設定して、局所的目標別に行為－価値関数を設計することを提案する。また、行為－価値関数は TD 法を用いて学習実験を行った。

最後に、局所的目標別に行為－価値関数をもつエージェントと局面全体で 1 つの行為－価値関数をもつエージェントの対戦実験を行った。その結果、局所的目標別に行為－価値関数をもつエージェントは 6 割以上の勝率をおさめることができた。

Design of an Action Value Function of various Local Goal in RoboCup Soccer

Tatsuya IRISAWA, Nobuo INUI, Yoshiyuki KOTANI

e-mail:irisawa@fairy.ei.tuat.ac.jp, kotani@cc.tuat.ac.jp

Tokyo Univ. of Agr and Tech, 2-24-16 Nakamachi, Koganei, Tokyo, JAPAN

Abstract

Fast decision of effective action is one of hard problem for RoboCup soccer agents. In this paper, we make action value function and resolve this problem. But it is too difficult to make effective action value function for all situations. So we set local goal by ball location and propose to make action value function various local goal. We examined the learning the weights of action value function using Temporal Difference learning.

We make the agents who have action value function various local goals and play a match against the agents who have only one action value function. The result shows the agents who have action value function various local goal win 60% game.

1 はじめに

近年、サッカーはマルチエージェント研究の標準問題として注目を集めている。サッカーは味方と協調することによって目標を達成する世界で最も有名なスポーツである。本稿では、電総研により開発されたRoboCupサッカーシミュレータを利用して実験を行っている[1]。サッカーシミュレータを利用する際、行動のメカニズムを作成しなければいけなく、それがチームの特徴になる。本研究の目的は、サッカーにおける有効な行動決定を学習することである。

サッカーにおける行動は、ボールを蹴ることができる位置とボールを蹴ることができない位置の 2 つのパターンに分けることができる。ボールを蹴ることができない位置での行動は、ボールを捕獲できる位置の予測[4] 、試合中に動的に変化するホームポジションの学習 [6] など様々な研究が行われている。ボールを蹴ることができる位置での行動も同様に研究がなされているが、それらは場面を限定しているため、実際の試合に適用できていない[2] [7]。そこで、本論文ではボールを蹴ることができる位置での行動決定の学習を行う。

本論文では、ボールを蹴ることができる位置では、行為－価値関数によって行動決定を行う。行為－価値関数とは現在の状態においてとりうる行動の価値を計算する関数のことである。行為－価値関数を用いると有効な行動を迅速に決定できるので、サッカーのような実時間性のある環境では有効な行動決定方法だといえる。しかし、局面全体で有効な行為－価値関数を設計するのは困難である。そこで、本論文では局所的目標別に行為－価値関数を設計することを提案する。

2 RoboCup サッカーにおける行動

RoboCup では行動はあらかじめ用意されてなく、開発者が独自のアルゴリズムで行動決定機構を作成する必要がある。そこで、本章では筆者が作成した行動決定機構を説明する。

サッカーにおいて行動はボールを蹴ることができるかどうかで大きく異なる。そこでまず、ボールを蹴ることができる位置での行動パターンとボールを蹴ることができない位置での行動パターンの 2 つにわけた。次にそれぞれのパターンでの行動と行動決定方法について述べる。

2.1 ボールを蹴ることができる位置での行動

2.1.1 行動の種類

ボールを蹴ることができる位置での行動には次のものを用意した。

・なにもしない
・シュート
・一番安全な味方へのパス
・一番敵ゴールに近い味方へのパス
・一番安全な空間へのパス
・クリア
・敵ゴールに向かってドリブル

安全な空間と味方は式(2.1)によって計算される。この式により敵エージェントから最も影響の少ない位置や味方を計算することができる。

$$\mathbf{P} = \underset{i}{\mathbf{argmax}} \sum_{j} \mathbf{d}_{ij} \qquad (2.1)$$

i ：候補ポジション、又は味方エージェント

$\mathbf{d}_{ij}$：ポジションiと敵エージェントjまでの距離

パス行動はそれぞれパスが届く範囲にいる味方エージェント、空間に制限して行う。
クリア行動はボールを敵ゴール方向に全力で蹴る行動である。

2.1.2 行動決定方法

ボールを蹴ることができる位置での行動は3章 で述べる行為－価値関数により決定する。

2.2 ボールを蹴ることができない位置での行動

ここでは、エージェントはまずボールを追いかけるかどうかを判断する必要がある。本論文におけるエージェントは味方エージェントの中で2番目までにボールに近い場合は無条件にボールを追いかけることにしている。

ボールを追いかけないエージェントは、ホームポジションを拠点に移動を行う。ホームポジションはサッカーには欠かせない概念である。エージェントには、ホームポジションをもとにその周辺に自分の領域が設定され、エージェントはその範囲内でのみ行動をとれるようになっている。また、ホームポジションは、状況に応じて変更することにより優れたチームプレイを実現することができる[6]。文献[6] では、遺伝アルゴリズムによりホームポジションを動的に変化させている。また文献[5] では簡単な計算式によってホームポジションを変化させている。本論文では、あらかじめ学習を行う必要のない文献[5] の方法によりホームポジションの変更を実現した。

3 行為－価値関数

行為－価値関数は現在の状態において取った行動の価値を計算する関数のことである。この価値はQ値と呼ばれている。行為－価値関数は、複数のパラメータを同時に扱うことができ、探索を行わずに行動決定ができるので、サッカーのような実時間性のある環境では有効な行動決定方法だといえる。

3.1 行為－価値関数による行動決定

本論文ではQ値は式(3.1)のような重み付き線形和で計算した。ここで i は行動の番号、j は状態の要素の番号、x_j は状態Xの j 番目の要素であり w_{ij} は重みである。

$$Q_i = \sum_j w_{ij} x_j \qquad (3.1)$$

Q値を用いると実際の行動は式(3.2) により決定される。

$$Action = \arg\max_i Q_i \qquad (3.2)$$

3.2 行為－価値関数の状態の要素

行為－価値関数の状態の要素として次のものを用いた。

・自分から敵ゴールまでの距離
・自分と敵ゴールのなす角度
・自分の安全度
・自分と最も安全な味方までの距離
・自分と最も安全な空間までの距離
・敵ゴール前のペナルティエリア内にいる敵の数
・最も安全な味方と敵ゴールまでの距離
・最も安全な味方と敵ゴールのなす角度
・最も安全な味方の安全度
・敵ゴールから最も近い味方と敵ゴールまでの距離
・敵ゴールから最も近い味方と敵ゴールのなす角度
・敵ゴールから最も近い味方の安全度
・定数1

ここで、安全度は式(2.1)によって計算される。

4 TD法による重みの学習

本論文では、行為－価値関数の重みの学習にTD法を用いる。TD法は予言間の差分を用いてパラメータを更新する学習法である。観測終了時には目標が達成できたかどうかに応じて報酬を与えている。TD法では近い未来の予言から学習を行うため効率よく学習を行うことができる。

TD法を用いると重みは次のように更新される。

$$W \leftarrow W + \sum_{k=1}^{T} \Delta W_k \qquad (4.1)$$

Wは重みの行列である(4.2)。mは行動の数、nは状態の要素数である。この行列では一行が一つの行動に対する重みになっている。

$$W = \begin{pmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{mn} & \cdots & w_{mn} \end{pmatrix} \qquad (4.2)$$

ΔW_t は重みの変化分であり次式によって計算される。

$$\Delta W_t = \alpha(P_{t+1} - P_t) \sum_{k=1}^{t} \lambda^{t-k} \mathrm{A}_k X_k \qquad (4.3)$$

α は学習率、λ は過去の観測状態に対する重みである。A_t は(4.4)で表す列ベクトルである。これにより、更新

される重みは時間tの時にとった行動の重みのみ更新させる。

$$A_t = \begin{pmatrix} a_t^1 \\ \cdot \\ \cdot \\ \cdot \\ a_t^m \end{pmatrix}, \quad a_t^i = \begin{cases} 1, & i = Action \quad \text{in time} \quad t \\ 0, & \text{otherwise} \end{cases} \tag{4.4}$$

X_tは状態ベクトルである。P_tは予想確率で式(4.5)によって表す。Tは観測終了時刻であり、S(Q)はQ値を1から0の値に変換するシグモイド関数である。

$$P_t = \begin{cases} S(\max Q_t), & t \neq T \\ 1, & t = T \quad \text{かつ} \quad \text{目標達成} \\ 0, & t = T \quad \text{かつ} \quad \text{目標失敗} \end{cases} \tag{4.5}$$

$$S(Q_t) = \frac{1}{1+e^{-Q}} \tag{4.6}$$

5 大局的目標と局所的目標

サッカーにおける大局的目標はゴールを決めることである。すべてのエージェントは大局的目標を達成させるための行動を行う。しかし、敵ゴールから遠く離れたエージェントがボールをもっている場合、ゴールを決めるための行動をとろうと考えるのは効率が悪い。このようなエージェントは単純にボールを前方の味方に渡すことを目標に行動をとればよい。なぜなら、サッカーではゴールを決めるためには敵ゴールに近づく必要があるので、前方の味方にボールを渡すことはゴールを決めるためには必ず必要になるからである。このことから、サッカーでは、ボールの位置により局所的目標があると考えることができる。

本論文では図1のようにボールの位置により局所的目標を設定した。図中の矢印は攻撃方向を示している。

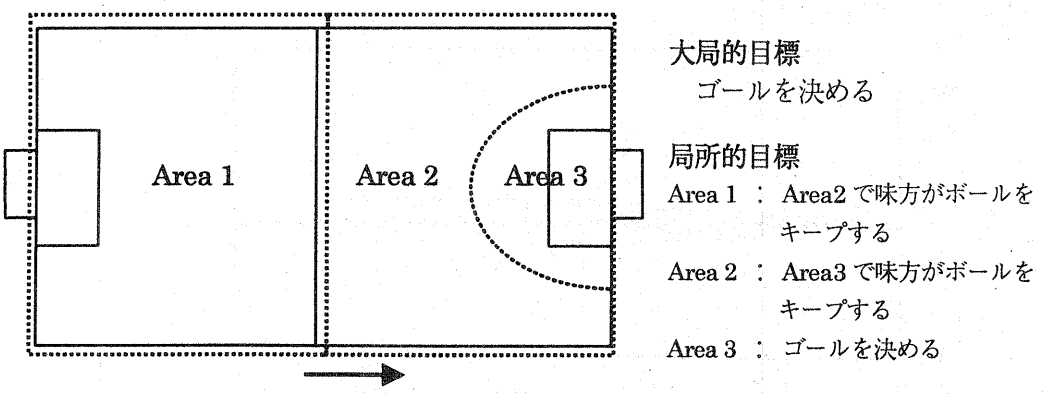

図1.フィールドの分割とAreaごとの局所的目標と大局的目標

6　行為－価値関数の重みの学習実験

ランダムプレーヤ同士に試合を行わせ、Area1、Area2、Area3、局面全体での訓練データを 10000 ずつ収集した。収集した訓練データを用いて、局所的目標別の行為－価値関数と局面全体での行為－価値関数の重みを学習する実験を行った。

6.1 訓練データの収集

訓練データは次のようにして収集した。

① 攻撃側エージェント(学習エージェント)11 人、守備側エージェント 11 人をフィールドに配置
② ボールを初期位置に配置
③ 試合開始
④ 終了条件になるまで試合を続ける
⑤ 観測終了までの状態とその時とった行動、目標が達成できたかどうかをファイルに書き込む

ボールの初期位置はそれぞれの Area にいる攻撃側エージェントがボールを蹴ることができる位置とした。試合の終了条件は表 1 のように設定した。また、すべての局面において、3000 ステップ経過した場合は目標失敗とした。

表 1.試合の終了条件

	Area1	Area2
目標達成	Area2で味方がボールをキープ	Area3で味方がボールをキープ
目標失敗	ゴールを決められる	Area1で敵がボールをキープ
	Area3	局面全体
目標達成	ゴールを決める	ゴールを決める
目標失敗	Area2で敵がボールをキープ	ゴールを決められる

6.2 学習の条件

学習率 α は初期値を 1000 とし学習が進むにつれて 0 に近づくようにした。λ は 0.95 とした。重みの初期値はすべて 0.5 とした。学習回数はそれぞれ 10000 回ずつ行った。

6.3 実験結果と考察

図 2,3 に重みの学習曲線を示す。

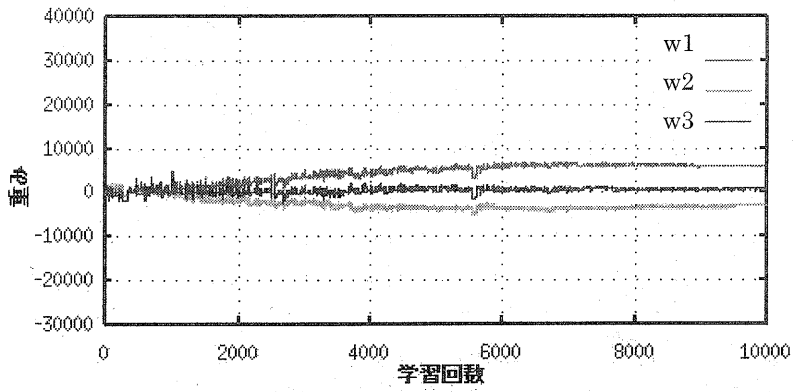

図 2.Area3 の学習曲線

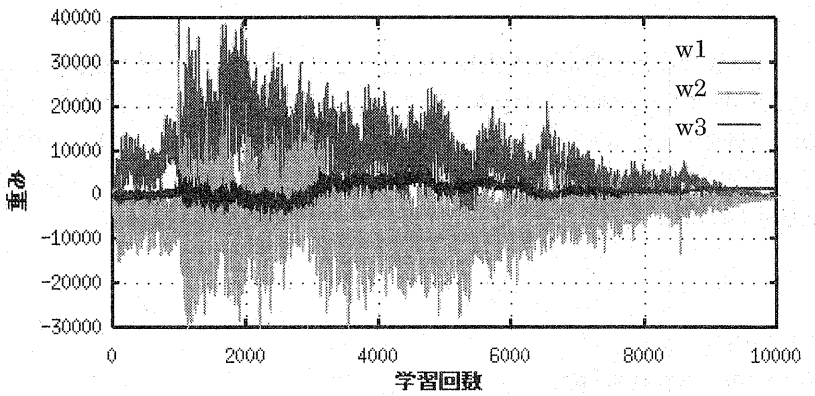

図 3.局面全体の学習曲線

w1、w2、w3 はそれぞれ自分から敵ゴールまでの距離、自分と敵ゴールのなす角度、自分の安全度のシュート行動に対する重みである

図 2、図 3 のどちらも 10000 回の学習により一定の値に収束している。このことから、Area ごとの学習も局面全体の学習もうまくいっていることがわかる。学習の様子を比較した場合、図 2 の振幅は図 3 に比べて緩やかである。これは図 2 では局面を敵ゴール前に狭めて学習を行っているため、局面全体での学習を行っている図 3 に比べて報酬を得るまでのステップ数が短いために 1 回の学習で更新する値が少ないからである。しかし図 2 の方が図 3 より収束するまでが速い。このことから局所的目標を用いたことによって有効な重みを見つけることが容易になったということがいえる。

6.4 対戦による評価

学習で得られた重みを用いて、Area ごとに行為－価値関数をもつエージェント(エージェント 1)と局面全体で 1 つの行為－価値関数をもつエージェント(エージェント 2)の対戦を行った。対戦は、実際の RoboCup のルールと同じ方法で 100 試合対戦を行った。結果は表 2 に示す。

表 2.対戦結果

	エージェント1	エージェント2
総得点	117	64
勝率	0.65	0.35

表 2 からわかるようにエージェント 1 が 6 割以上の勝率を収めていて、局所的目標別に行為－価値関数を設計することは有効だということがいえる。しかし、総得点を見ると、エージェント 1、2 ともに低く平均すると 1 試合に 1 点とれるかどうかである。これは対戦実験に用いたエージェントの基本技能が低いという原因もあるが、学習中の対戦相手にも問題がある。今回用いた対戦相手はランダムプレーヤである。ランダムプレーヤを用いて学習を行うと学習の偏りはなくなるが有効な行動を多く行うことはなくなる。そのため訓練データ数が少ないと有効な行動が選ばれない場合がある。そのためランダムプレーヤによる実験ではデータ数を多くする必要がある。今回集めた 10000 の訓練データではまだ不十分だったということが考えられる。

7 まとめ

本論文では、ボールの位置により局所的目標を設定し、局所的目標別に行為－価値関数を設計することを提案した。また、TD法を用いて行為－価値関数の重みを学習する実験を行った。そして、学習された重みの有効性を確かめるために、大局的を行ったエージェントと局所的目標により学習を行ったエージェントの対戦実験を行った。その結果、局所的目標により学習を行ったエージェントは大局的目標により学習を行ったエージェントに対し6割以上の勝率を収めることができた。

参考文献

[1] RoboCup web page,
http://www.robocup.org/02.html

[2] NODA Itsuki, MATSUBARA Hitoshi and HIRAKI Kazuo, Learning Cooperative Behavior in Multi-agent Environment, PRICAI'96, pp.570-579, 1996.

[3] Peter Stone, Manuela Veloso , and Patrick piley : The CMUnited-98 Champion Simulator Team,RoboCup98:Robot Soccer World Cup Ⅱ, Springer Verlag, Berlin, 1999.

[4] Manuela Veloso, Peter Stone, and Michael Bowling, Anticipation: A Key for Collaboration in a Team of Agents, ICAA, 1998.

[5] 遠藤 和昭 ：目標の階層化による RoboCup エージェントの行動決定, 東京農工大学修士論文, 2000.

[6] 村田 哲也, 山本 雅人, 鈴木 恵二, 大内 東：GA によるサッカーエージェントの動的配置探索問題に関する研究, 人工知能学会誌, Vol14 No3 pp456-464, 1999.

[7] 秋山 直之, 鈴木 恵二, 山本 雅人, 大内 東：サッカーエージェントによる協調パスプレイの生成, ゲーム・プログラミングワークショップ '99,pp193-199, 1999.

[8] 薄井 克俊, 鈴木 豪, 小谷 善行：TD法を用いた将棋の評価関数の学習, ゲーム・プログラミングワークショップ '99, pp31-38, 1999.