2ZE-06

TD 法によるサッカーエージェントの行動の学習

入沢 達矢, 遠藤 和昭, 乾 伸雄, 小谷 善行
(東京農工大学 工学部 電子情報工学科)

1 はじめに

実時間性をもったマルチエージェント環境においてエージェントが有効な行動を一意に決定することは困難である。それは、エージェントの行動は他エージェントの行動に依存するが、時間が限られているのでそれらを十分考慮に入れる時間がないからである。そこで我々はサッカーにおける行動決定に行為－価値関数を利用する。

本研究ではロボカップサッカーを題材に行為－価値関数によって行動を決定するエージェントを作成した。そして、TD 法を用いて行為－価値関数のパラメータを学習させ、行動の性能向上を計る実験を行った。

2 行為－価値関数による行動決定

2.1 行為－価値関数

行為－価値関数とは、現在の状態においてある行為を行ったときのその行為の価値を計算する関数である。また、ここで計算された価値のことを Q 値という。状態 X において行動 a_i を行ったときの Q 値は次式で計算される。

$$Q(a_i, X) = \sum_j w_{ij} x_j$$

w_{ij} ：重み

x_j ：状態の j 番目の要素

そして、エージェントはすべての行動 a_i に対してこの式を適用して Q 値が最大になる行動をとる。

2.2 行為－価値関数のサッカーへの適用

サッカーでは、ボールとエージェントとの位置関係によって行動の種類が大きくことなる。そのうち、エージェントがボールを蹴ることができる位置にいる場合には、エージェントが選択できる行動として、シュート、ドリブル、パス、クリアといった多彩な行動が存在する。そこで、我々は、ボールを蹴ることができる位置にいる場合において有効な行動を選択させるために、行為－価値関数を利用する。

2.3 状態

本研究では、行為－価値関数の状態の要素として次のものを扱う。

x_0：自分から敵ゴールまでの距離
x_1：自分の周りの密集度
x_2：自分からいちばん近い敵までの距離
x_3：自分からいちばん近い味方の密集度
x_4：自分と敵ゴールのなす角度
x_5：自分と敵ゴールの間にいるエージェント数

ただし、密集度は自分の周り半径 10m 以内にいるエージェント数によって計算する。

2.4 ボールを蹴ることができる位置での行動

本研究では、ボールを蹴ることができる位置での行動として次のものを扱う。

a_0：なにもしない
a_1：シュート(敵ゴールに向かって蹴る)
a_2：パス(味方エージェントに向かって蹴る)
a_3：クリア（前方に大きく蹴る）
a_4：ドリブル

2.5 TD 法による重みの更新

TD 法とは、過去の観測状態と最終結果から学習する手法である。TD 法では次式によって重みの更新を行う。

$$W \leftarrow W + \sum_{k=1}^{T} \Delta X_k$$

$$\Delta X_t = \alpha(P_{t+1} - P_t) \sum_{k=1}^{t} \lambda^{t-k} X_t$$

Learning Behavior of Soccer Agent by Temporal Differences
Tatsuya IRISAWA, Kazuaki ENDOU,
Nobuo INUI, Yoshiyuki KOTANI
Department of Computer Science,
Tokyo University of Agriculture and Technology
2-24-16 Nakamachi Koganeishi Tokyo Japan

T ：観測終了時刻
α ：学習率
λ ：過去の状態に対する重み
P：予想確率

ただし、予想確率PはQ値を次のシグモイド関数によって0～1の予想確率に変換したものである。

$$P(Q)=\frac{1}{1+e^{\frac{-Q}{5000}}}$$

また、観測終了時刻では、結果により1か0の報酬を与えることにより学習させる。

3 実験

3.1 実験方法

実験はゴール前に、攻撃側2、守備側1の2対1の場面を作成し、攻撃側のエージェントのうち1人に対して学習を行う。そして、学習のための報酬は、

- ゴールを決めたら報酬1
- 10秒たってもゴールがきまらなかったら報酬0

とした。重みの初期値は0.5、αは0.1、λは0.8として学習を行う。学習回数は10000回とした。

3.2 実験結果

図1に学習経過におけるゴール成功率の推移を示す。

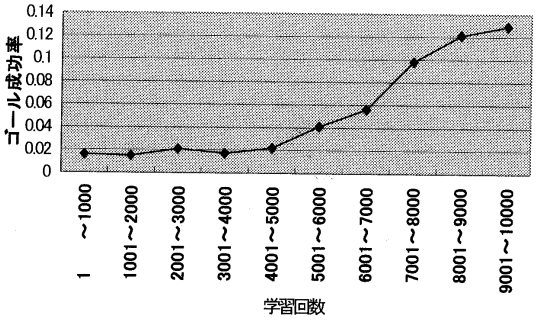

図1.ゴール成功率の推移

図2に学習経過における行動ごとの選択された割合の推移を示す。

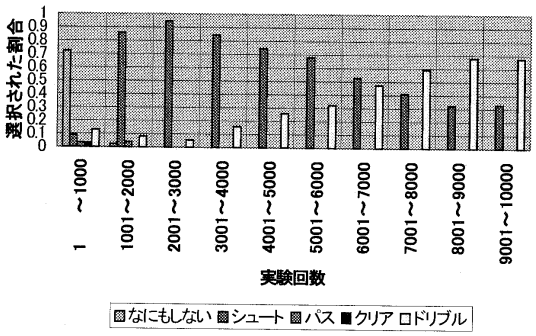

図2.行動の選択された割合の推移

また、この実験から、最終的にゴールまである程度近づいたあとシュートをうつという行動が生成された。

4 考察

- ゴールを成功させることができるようになったことから学習されているといえる。
- 図1、2から、ゴール前2対1の場面でクリア、パスといった行動は有効でないといえる。
- 図2から最初の2000回はゴール前においての有効な行動を探しているといえる。
- 図2から2000回以降は、シュートをうつタイミングを学習しているといえる。

5 おわりに

本稿では、TD法によるサッカーエージェントにおける行動の学習について述べた。その結果、次のことがわかった。

- サッカーのような実時間性をもったマルチエージェント環境において行為－価値関数を利用することは有効である。
- 行為―価値関数をTD法を用いて学習させることは有効である。

参考文献

[1] 薄井 克俊：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, pp.31-38, 1999

[2] Stuart Russell, Peter Norvig 著, 古川 康一 監訳：エージェントアプローチ人工知能 20章 強化学習, 共立出版, pp.601-628, 1997