サッカーロボットにおける振舞の認識と行動

Understanding Behaviors of The Other Agent

Applying Reinforcement Learning to Soccer Robot in Multi-agent Environment

○ 正　浅田　稔　(阪大)　　准　内部　英治　(阪大)
准　細田　耕　(阪大)

Minoru ASADA, Osaka University, 2-1, Yamadaoka, Suita, Osaka
Eiji UCHIBE, Osaka University
Koh HOSODA, Osaka University

Abstract : We have previously applied the vision-based reinforcement learning for the integration of a set of tasks of which states are not completely independent of each other in an environment including the other competitive agent. However in a multi-agent environment, it is difficult for multiple robots to learn simultaneously because the environment including the other learning agent changes randomly in the early learning state. In this paper, we discuss the issues in the understanding behaviors of other agent from observation. Principal component analysis is used to analyze the observation matrix, and the robot computes the loss of information in order to discriminate the behavior between 4 typical behaviors. Simulation results are shown and the discussion is given.

Key Words : behavior understanding, multi-agent environment, observation, soccer robot

1　はじめに

近年，合目的な行動を獲得するための手法として，強化学習が注目されている．これまで，強化学習を用いて，他のエージェントが存在する環境で，複数のタスクを達成する手法について報告してきた [5].

我々の目標は，複数のエージェントが存在する環境で，それぞれのエージェントが互いに競合・協調といった関係のあるタスクを与えられた場合に，効率の良い学習を行なう枠組を提案することである．しかし，複数のエージェントが同時に学習する場合，単一のエージェントのみが学習する場合と比較して，学習結果が悪くなることが指摘されている [1,2]．また，環境が非同期に変化するといった問題もある [3].

この原因として，同時に学習を行なう場合，他のエージェントの行動が固定した政策に基づいておらず，結果として，本来学習すべき環境と異なる環境で学習を行なっている点が挙げられる．この問題を回避するために，エージェント間で学習に関する情報交換を行ない，順序を決めて学習を行なう方法が考えられる．自分以外のエージェントの行動を観測し，学習するエージェントを一つ決定することで，全体としてのエージェント群の能力は，次第に改善されると考えられる．

そこで本報告では，観察によって他のエージェントの行動を理解 (同定) するための一手法を提案する．他のエージェントの行動を理解することは，強化学習を適用する判断材料になる．さらに，学習時の効果的な行動戦略にも利用できると考えられる．

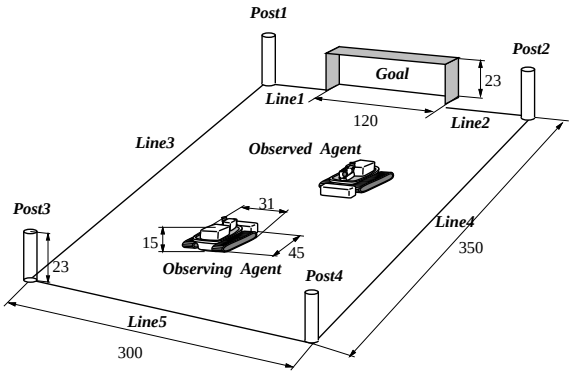

Fig.1 Environment

2　問題設定

2.1　タスクと仮定

設定として，**Fig. 1** のような環境 (ゴール，ポスト4本，ライン5本) を考える．観測者以外に，もう一体

のエージェントが存在し，そのエージェントの行動を理解することを目的とする．

エージェントは同一のものを使用し，PWS(Power Wheeled Steering) を採用する．また，自身の幾何学的パラメータや動的特性などに関する知識はもってないと仮定する．

エージェントにはカメラが搭載されており，このカメラだけから，環境の情報を獲得しなければならない．具体的には，画像上での位置の x 座標，大きさなどを検出する．エージェントに関しては，向きも検出できると仮定する．

2.2 観測されるエージェントの行動戦略

極論すると，エージェントの行動戦略は，(a)「環境内に存在する物体だけに依存」，(b)「自身の内部状態だけに依存」の 2 種類あり，実際の戦略はこの 2 種類の戦略の中間をとっていると考えられる．

そこで観測される行動戦略として，

(i) 静止している，

(ii) ゴールに向かう，

(iii) 観測エージェントに向かう，

(iv) ランダムウォーク，

といった 4 種類を準備する．実際の行動は，環境からの制約を受けるため，必ずしも「自分自身の内部状態だけに依存」するわけではないが，(iv) は (b) の一例と考えられる．また，(ii),(iii) は (a) のなかで，対象物が静的が動的かの違いがある．(i) は 他のエージェントが単純に静止環境の一部とみなせる例である．

ここで，一試行の観測中に，観測されるエージェントの行動戦略は，上記のいずれか一つだけを使用しているものとし，途中で変更しないと仮定する．

2.3 観測するエージェントの行動戦略

他のエージェントの行動 (状態遷移) を理解するためには，そのエージェントを観察しなければならない．しかし，観測エージェントがランダムに行動していたのでは，効率が悪い．つまり，観察のための行動戦略として，相手エージェントを注視し続けるような戦略が必要となる．

環境に関する先見的な知識を持たない，という仮定からすれば，最終的には，強化学習の一つである Q 学習などを用いてリアクティブな行動戦略を獲得しなければならないが，ここでは問題の簡単化のため，単純な制御則を構成し，それによって追跡を行なうものとする．

3 多変量データ解析に基づく行動理解

3.1 データの収集

2.3 節で述べた行動戦略に従って，観測エージェントは環境の状態を観測する．観測されたデータは

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \tag{1}$$

として，表現される．ここで，n は時刻，m は説明変数の個数を表す．説明変数には，自分自身の行動と相手エージェントに関する状態量と，同時に観測されたその他の環境中の対象物に関する状態量を用いる．説明変数間で単位が異なるため，データ行列を列標準化 (平均0，分散1) し，それを $\boldsymbol{Z}$ とする．

具体例 相手エージェントを観測したとき，同時にゴールが見えていたとする．そのときの説明変数として，観測エージェントの行動と，相手エージェントの画像上での位置の x 座標，高さ，向き，そしてそれぞれの変化量，さらに，ゴールの位置の x 座標，高さ，向き，その変化量といった合計 $2+3\times2+3\times2=14$ 個の説明変数を準備する．そのときの説明変数として，相手エージェントの画像上での位置の x 座標，高さ，向き，さらに，ゴールの位置の x 座標，高さ，向きといった合計 $3+3=6$ 個の説明変数を準備する．

3.2 観測データの解析

獲得された行列に，主成分分析 [4] を適用する．実際には，説明変数間で単位が統一されていないため，各列の要素は，平均0，分散1となるように標準化した行列 $\boldsymbol{Z}$

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{D}_S^{-1/2}, \tag{2}$$

$$\boldsymbol{D}_S^{-1/2} = \mathrm{diag}\,[1/s_{x1}\, 1/s_{x2} \cdots 1/s_{xm}] \tag{3}$$

を用いる．ここで s_{xj} は説明変数 X_j の標準偏差である．主成分分析により，非負の固有値 λ_i とそれに対する固有ベクトル $\boldsymbol{a}_i$ $(i=1\cdots m)$ が求まる．これによって第 i 主成分 f_i が計算される．f_i は

$$f_i = \boldsymbol{a}_1^T \boldsymbol{z} = a_{1i}x_{*1} + a_{2i}x_{*2} + \cdots + a_{mi}x_{*m} \tag{4}$$

で与えられる．これがエージェントに関する状態と，環境の状態に関する状態の関係を表している．関係を良く表している主成分を用いて，行動理解に利用する．全ての主成分を利用するのは，現象を簡単化して理解するという観点からは望ましくない．そこで，ここでは第 j 主成分までの累積寄与率が

$$\frac{\sum_{i=1}^{j}\lambda_i}{\sum_{i=1}^{m}\lambda_i} \geq 0.9 \tag{5}$$

であるときの主成分 $f_1\cdots f_j$ を採用する．これは，第 j 成分までの主成分を用いて，データの 90% 以上が説明できることを意味している．

3.3 行動戦略の判別

行動を判別するために，前節で獲得された主成分を利用する．今，それぞれの行動 $i\,(i=1\cdots4)$ に対して，観測データ行列 $\boldsymbol{Z}_i$ から p_i 個の主成分 $f_k^i\,(k=1\cdots p_i)$ が獲得されたとする．

次に，新しい $n'\times m$ の観測データ行列 $\boldsymbol{Z}'$ が獲得されたとする．このとき，時刻 t での，それぞれの行動 i に対する主成分を用いた情報損失量 $loss^i(t)$ は，

$$loss^i(t) = \sqrt{\sum_{l=1}^{m}(z'_{tl})^2 - \sum_{j=1}^{p_i}(f_j^i)^2} \tag{6}$$

と計算される．情報損失量がもっとも小さくなる $i_{\min}$ を $\boldsymbol{Z}'$ が表現している行動であるとする．例として，$m = 2$，$p_i = 1$，行動戦略が2種類(A,B)の場合を示す．この場合，情報損失量は戦略Bの方が小さい．

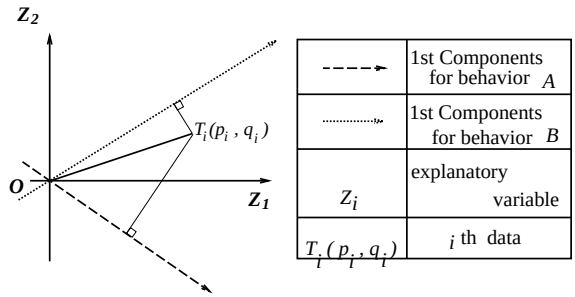

Fig.2 discrimination

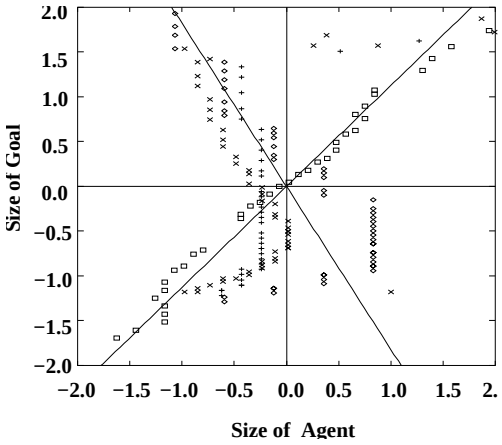

Fig.4 toward goal(3)

4 シミュレーション結果

2.2 節で述べた 4 種類の行動の理解をシミュレーションによって行なった．説明変数の候補として，それぞれ

- 自分の行動:2種類 (前進速度 v と 角速度 ω)
- 相手エージェント:3種類 (画像上での位置の X 座標，大きさ，向き)
- ゴール:3種類 (相手エージェントと同様)
- ポスト:2種類 (画像上での位置の X 座標と大きさ)
- ライン:2種類 (画像上での位置の Y 座標と傾き)

を準備し，この組合せでデータ行列の説明変数は決定される．例えば，相手エージェントとゴールだけが見えている場合には，説明変数の個数は自分の行動も含めて合計 $2 + 3 + 3 = 9$ 個となる．

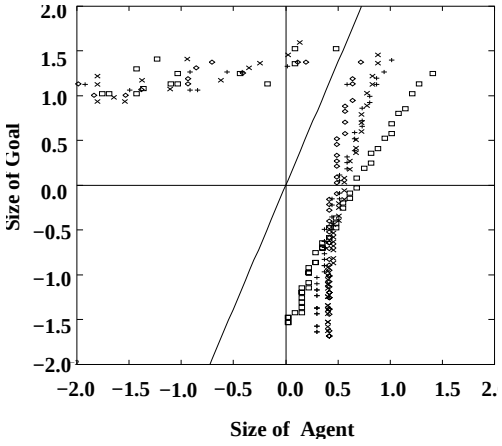

Fig.5 toward agent(3)

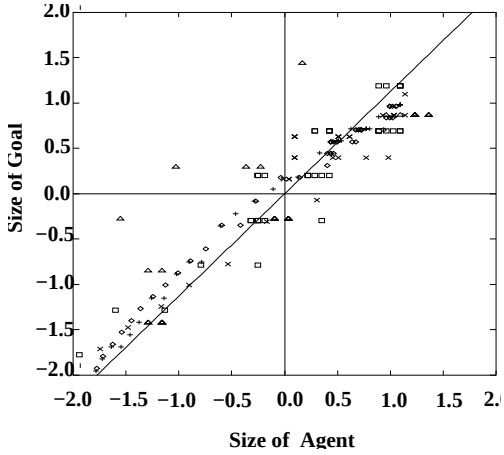

Fig.3 stationary(3)

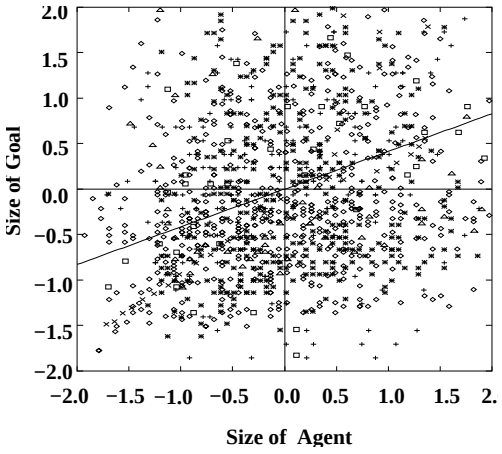

Fig.6 random walk(8)

Fig.3〜6 に，それぞれの行動について獲得された，相手エージェントとゴールに関するデータ行列を列標準化したものの一部を示す．データの可視化のため，X 軸を画像上でのエージェントの大きさ，Y 軸を画像上でのゴールの大きさとする．括弧内の数字は，式 (5) で決定された主成分の個数である．

(d) の場合には，画像の解像度の低さのため，標準化したデータの分布には，疎密が見られるが，ほぼ一様に観測データが分布している．他のそれぞれの行動には，

1. (a) 相手が大きくなると，ゴールも大きくなる．
2. (b),(c) 相手の大きさは一定であり，ゴールの大きさは変化する．

といった特徴がある．(b),(c) は，エージェントの向きの軸により特徴づけることができる．主成分分析により，これらのデータを説明できるような，新しい軸 (主成分) が発見できる．このような関係が，ポストやラインなどに対しても求められる．

次に，実際にゴールへ向かう行動戦略を主成分を用いて判別した結果を **Fig.8** に示す．ここで，Y 軸は情報損失量を表す．相手が静止している場合には，主成分がかなり異なるため，結果として情報損失量が他と比べて，非常に大きな値となり，Fig.8には示していない．観測の初期段階で，自分自身に向かう行動と識別を失敗しているが，だいたいの場所で正確に識別できた．

失敗の原因として，ゴールへ向かう行動戦略が複数の戦略から構成されていることが挙げられる．ゴールへ向かう行動戦略は 1. ゴールの探索，2. ゴールまでのナビゲーション の2つの戦略に分解できる (**Fig.7**参照)．3.2節の方法で獲得した主成分は，2. の方の戦略

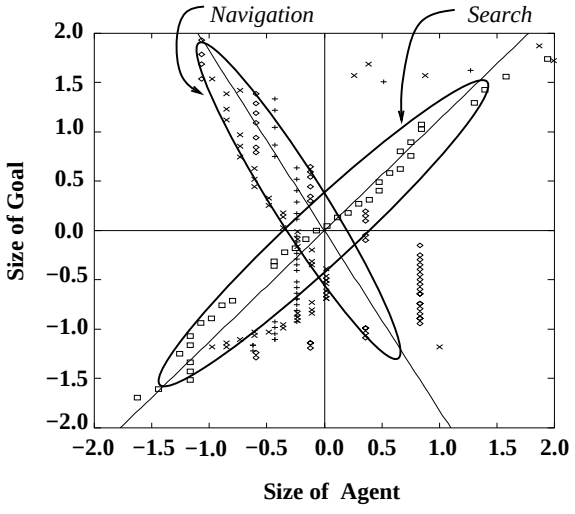

Fig.7 Analysis of toward-goal behavior

を良く説明しているのに対し，観測の初期段階では，

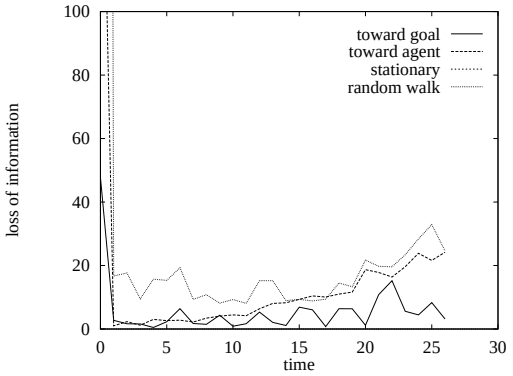

Fig.8 Result of discrimination

1. の行動戦略をとっている．結果として，この間の情報損失量は増加することになる．

5 おわりに

本稿では，マルチエージェント環境において，強化学習を適用するために，他のエージェントの行動を理解する手法を提案し，シミュレーションによってその有効性を検証した．

今後の方針について，まず，実機を用いて，提案した手法の有効性を検証することが挙げられる．データの分析に関しては，獲得されたデータ行列 $\boldsymbol{X}$ が時系列データであることを考慮した分析を行なう必要がある．

また，ボールのような自分自身では行動できない，受動的なエージェントを環境に組み込むことが考えられる．

参考文献

[1] J. A. Boyan. Modular Neural Networks for Learning Context-Dependent Games Strategies. Master's thesis, University of Chicago, 1991.

[2] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proc. of the 11th International Conference on Machine Learning*, pp. 157–163, 1994.

[3] 溝口. 行動メディア. 第12回日本ロボット学会学術講演会予稿集, pp. 843–844, 1994.

[4] 柳井. 多変量データ解析—理論と応用—. 行動計量学シリーズ 8. 朝倉書店, 1994.

[5] 内部, 浅田, 野田, 細田. 視覚に基づく強化学習による移動ロボットの多重タスクの 達成. 第12回日本ロボット学会学術講演会予稿集, pp. 609–610, 1994.