他のエージェントの行動理解

—サッカーロボットにおける強化学習のマルチエージェント環境への適用に向けて—

◯内部 英治　　浅田 稔　　細田 耕

大阪大学工学部

Understanding Behaviors of Other Agents

—Applying Reinforcement Learning to Soccer Robot in Multi-agent Environment—

◯Eiji UCHIBE　　Minoru ASADA　　Koh HOSODA

Osaka University

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[2)]．

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)]．

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．

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．

2 問題設定

2.1 タスクと仮定

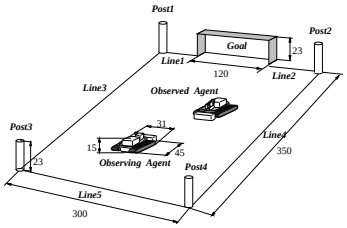


Fig.1 Environment

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,「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を理解するためには，ランダムに行動するよりも，相手を注視していた方が，データを獲得しやすい．最終的には，学習によって，注視行動を獲得すべきであるが，ここでは，行動理解に焦点をしぼるために，対象物の画像上の重心位置を，中心にする制御則を構成した．

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．

3.2 観測データの解析

獲得された行列に，主成分分析[3)] を適用する．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．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．ここでは第 j 主成分までの累積寄与率が

$$\frac{\sum_{i=1}^{j} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \geq 0.9 \tag{2}$$

であるときの主成分 $f_1 \cdots f_j$ を採用する．

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 $i\,(i = 1 \cdots 4)$ 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，

$$loss^i(t) = \sqrt{\sum_{l=1}^{m}(z'_{tl})^2 - \sum_{j=1}^{p_i}(f_j^i)^2} \qquad (3)$$

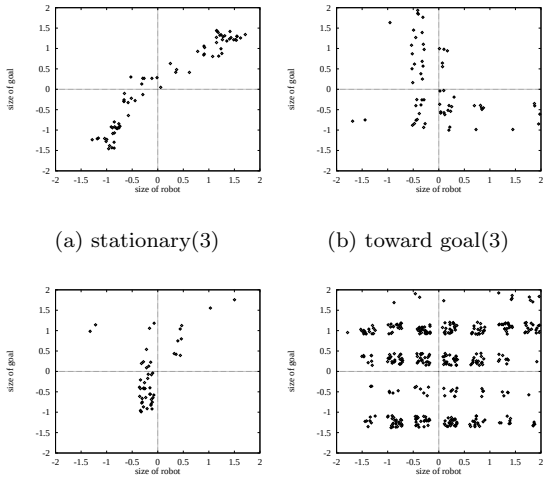
と計算される．情報損失量がもっとも小さくなる $i_{\min}$ を $\boldsymbol{Z}'$ が表現している行動であるとする．

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．



(a) stationary(3) (b) toward goal(3)

(c) toward agent(3) (d) random walk(8)

Fig.2 Distribution of acquired data

Fig.2 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 (2) 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.3** 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.3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．3.2節の方法で獲得した主成分は，2. の方の戦略を良く説明しているのに対し，観測の初期段階では，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．

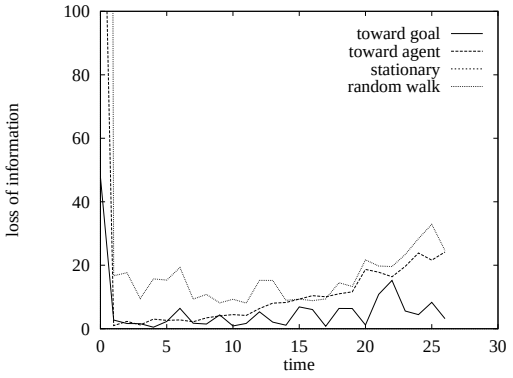


Fig.3 Result of discrimination

参考文献

[1] 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.

[2] 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.

[3] 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.