マルチエージェント環境における部分空間同定法を用いた状態空間の構成と行動獲得

◯内部 英治　　浅田 稔　　細田 耕
大阪大学大学院 工学研究科

State Space Construction and Behavior Acquisition Using Subspace Identification in Multi-Agent Environments

◯Eiji Uchibe　　Minoru Asada　　Koh Hosoda
Osaka University

Abstract — *This paper proposes a method which estimates the relationship between the learner's behaviors and the other agents' ones through interactions (observation and action) using the method of system identification to construct a state space in such an environment. Next, reinforcement learning based on the estimated state vectors is utilized to obtain the optimal behavior. The proposed method is applied to soccer playing physical agents, which learn to cope with a rolling ball and moving other agents.*

1　はじめに

ロボットに自律的に目的行動を獲得させる手法として，強化学習法が注目されている．しかし，マルチエージェント環境では，学習者にとって他者の行動がランダムに観測される場合があるため，通常の強化学習をそのまま適用するのは困難である．学習を成功させるためには，学習者は他者の行動を自分自身の観測と行動を通して予測できる必要がある．Peter and Veloso[2] や Uchibe et al.[3] は複数台のロボットが存在する環境での学習の問題を扱っているが，センサ出力を状態として利用しており，センサの変化量と行動が 1 対 1 に対応しない場合は学習は困難になる..

そこで本報告では，学習者の観測と行動を通して，学習者と他者の行動の関係を局所予測モデルとして推定し，局所予測モデルと強化学習を統合する手法を提案する．

提案する手法を簡単なサッカーゲームに適用する．タスクは他のロボットがパスしてきたボールをゴールにシュートするタスクである．環境は静的エージェント (ゴール)，受動エージェント (ボール)，能動エージェント (移動ロボット) から構成され，学習者はそれぞれに対して局所予測モデルを構築する．学習者は予測モデルを構築した後に，強化学習によって目的行動の学習を開始する．実ロボットによる実験結果を示し，本手法の有効性を検証する．

2　観測と行動による内部モデルの構築

2.1　局所予測モデル

学習が成功するためには，学習者が継続して起こる状態を予測できる必要がある．しかし，ロボットはセンシング能力の限界から，予測に必要な情報を完全に知ることはできない．

そこで，学習者ができることは学習者自身の行動が他者に与える影響を学習者自身の観測を通して解析することだけであると考えられる．以下では，学習者の行動と他者の行動の関係を，学習者の入出力のシーケンスから構築することを述べる．

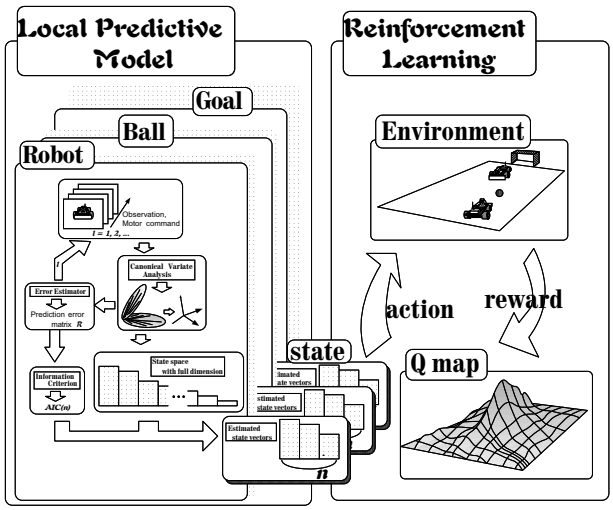


Fig.1 the flow of the proposed method

Fig.1 は局所予測モデルの処理の流れを示している．

2.2　正準変量解析 (CVA)

多入力多出力のシステムを同定する手法として，部分空間同定法がある．そこで，局所予測モデルを構築するために，部分空間同定法の一つである正準変量解析 (CVA)[1] を適用する．

いま，局所予測モデルが

$$\begin{aligned} \boldsymbol{x}(t+1) &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t) + \boldsymbol{w}(t), \\ \boldsymbol{y}(t) &= \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{D}\boldsymbol{u}(t) + \boldsymbol{v}(t), \end{aligned} \quad (1)$$

と表現できると仮定する．ここで $\boldsymbol{u}(t) \in \Re^m$，$\boldsymbol{y}(t) \in \Re^q$ は入出力ベクトル，$\boldsymbol{x}(t) \in \Re^n$ は状態ベクトル，$\boldsymbol{v}(t) \in \Re^q$，$\boldsymbol{w}(t) \in \Re^n$ は白色雑音ベクトルである．CVA は式 (1) を正則行列 $\boldsymbol{T}$ によって変換した状態空間モデル

$$\begin{bmatrix} \boldsymbol{\mu}(t+1) \\ \boldsymbol{y}(t) \end{bmatrix} = \boldsymbol{\Theta} \begin{bmatrix} \boldsymbol{\mu}(t) \\ \boldsymbol{u}(t) \end{bmatrix} + \begin{bmatrix} \boldsymbol{T}^{-1}\boldsymbol{w}(t) \\ \boldsymbol{v}(t), \end{bmatrix}, \quad (2)$$

を推定する．ここで $\boldsymbol{\mu}(t)$ は推定された状態ベクトルである．アルゴリズムの詳細については文献 [1] を参照されたい．

厳密には，学習者は環境全体に対して局所予測モデルを構築する必要があるが，次元の増加のため，必要なデータを収集することが困難になる．よって，ここでは学習者は画像処理等によって抽出された対象に関する特徴量ごとに局所予測モデルを構築する．

2.3 状態ベクトルの次元の決定

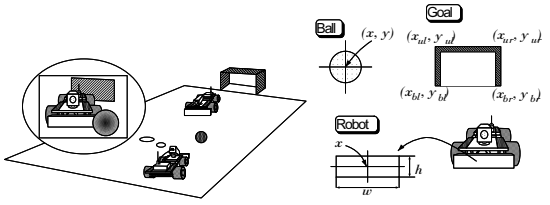
状態ベクトルの次元 n と考慮する履歴長さ l を決定するのは重要な問題である．n を決定するために，赤池の情報量規準 (AIC) を適用する．AIC を最小にする n を最適な次元とする．

3 強化学習

局所予測モデルを構築した後に，学習者はその領域内で強化学習を適用する．ここでは Q 学習 [4] を用いる．Q 学習と局所予測モデルを統合するために，推定された n 次元状態ベクトル $\boldsymbol{\mu}$ を離散化する必要がある．いま，状態ベクトル $\boldsymbol{\mu}$ は共分散行列は単位行列に規格化されているため，今回はそれぞれの状態を閾値 $\pm\theta$ で 3 つに離散化する．

4 実験結果

提案する手法を **Fig.2(a)** に示す簡単なサッカーゲームに適用する．各ロボットは一つのカメラを中心部に



(a) The task　　(b) Image feature points

Fig.2 The environment and our mobile robot

搭載した，4WS システムの移動台車である．モータコマンドとして駆動速度 v とステアリング角 ϕ をそれぞれ 3 通りに離散化した．また，各ロボットが観測する画像特徴量を **Fig.2(b)** に示す．

報酬はシュート，パスが成功したときに正の報酬を，衝突したときに負の報酬を与えた．学習するのは 1 台だけで，学習していないロボットは政策を固定し，適当な段階で学習するロボットを交替させた．

Table 1 は，提案手法と画像特徴量を状態として強化学習をしたときの比較結果である．画像特徴量だけを用いた場合，ボールの転がりなどのため，獲得される行動価値関数が局所解を持つことになり，タスクを達成できない．

Table2 は実環境でデータを収集，学習したときの同定結果を示している．ただし，計算機上で獲得した学習結果を実ロボットに移植し，実環境での学習時間

Table 1 Performance result

state vector	success of shooting (%)	success of passing (%)
current features	10.2	9.8
proposed method	78.5	53.2

Table 2 The estimated dimension

	from the shooter				from the passer			
	l	n	$\log \lvert\boldsymbol{R}\rvert$	AIC	l	n	$\log \lvert\boldsymbol{R}\rvert$	AIC
ball	4	4	1.88	284	4	4	1.36	173
goal	1	3	-1.73	-817	$*$	$*$	$*$	
robot	5	4	3.43	329	5	4	2.17	284

を短縮した．ここで，状態の次元が同一であっても，その表現は各ロボット間でまったく異なる点に注意されたい．これは個体差や経験の違いによるもので，ロボット間での学習結果の伝播については今後の方針である．

最後に，獲得された行動例を **Fig.3** に示す．

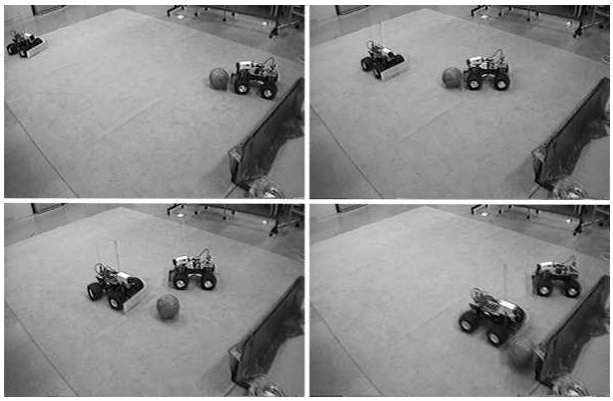


Fig.3 Acquired behavior

5 おわりに

本報告では，他者が存在する環境に強化学習を適用し，行動獲得するための手法を提案し，その有効性を検証した．

参考文献

1) W. E. Larimore. Canonical variate analysis in identification, filtering, and adaptive control. In *Proc. 29th IEEE Conference on Decision and Control*, pp. 596–604, Honolulu, Hawaii, December 1990.

2) P. Stone and M. Veloso. Using machine learning in the soccer server. In *Proc. of IROS-96 Workshop on Robocup*, 1996.

3) E. Uchibe, M. Asada, and K. Hosoda. Behavior coordination for a mobile robot using modular reinforcement learning. In *Proc. of the 1996 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1329–1336, 1996.

4) C. J. C. H. Watkins and P. Dayan. Technical note: Q-learning. *Machine Learning*, pp. 279–292, 1992.