マルチエージェント環境における行動学習のための部分空間同定法による状態空間の構成

内部 英治　浅田 稔　細田 耕

uchibe@er.ams.eng.osaka-u.ac.jp
大阪大学大学院 工学研究科 知能機能創成工学専攻
〒565-0871 大阪府吹田市山田丘2-1

学習者以外に能動的に行動できるエージェントが存在する環境では，学習者がエージェントを含めた環境の変化を予測できない限り，適切な行動を学習によって獲得できない．本報告では，学習者の行動が他のエージェントに及ぼす影響を観測を通して推定することにより，エージェントの分類および行動戦略を識別する方法を提案する．エージェントのモデルを同定するために，部分空間同定法の一つである正準相関分析に対し，赤池の情報量規準を応用する．モデルを獲得した後に，得られた状態ベクトルに基づき強化学習を適用する．提案する手法をサッカーロボットに適用し，本手法の有効性を検証する．

State Space Construction based on Subspace Identification for Behavior Acquisition in Multi Agent Environments

Eiji Uchibe, Minoru Asada, and Koh Hosoda

Graduate School of Eng., Dept. of Adaptive Machine Systems
Osaka University, 2-1, Yamadaoka, Suita, Osaka 565-0871, Japan

This paper proposes a method that acquires the purposive behaviors based on the estimation of the state vectors. In order to acquire the cooperative behaviors in multi robots environments, each learning robot estimates the local predictive model between the learner and the other objects separately. Based on the local predictive models, robots learn the desired behaviors using reinforcement learning. The proposed method is applied to a soccer playing situation, where a rolling ball and other moving robots are well modeled and the learner's behaviors are successfully acquired by the method. Computer simulations and real experiments are shown and a discussion is given.

1　はじめに

実世界で与えられたタスクを遂行することを自律的に獲得できるロボットを実現することは，ロボティクスと AI の中心課題の一つである．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，学習者の行動が，必ずしも自分自身の観測と 1 対 1 には対応しないため，通常の強化学習をそのまま適用することは困難である．マルチエージェント環境での学習を困難にしている理由として，

A 他者の行動政策は，学習者にとって未知であり，センサから得られる瞬間の情報だけでは，次の状況を予測することは困難である．

B 特に学習の初期段階において，他者のランダムな行動戦略が，学習者の学習過程に悪影響を及ぼす．

といったことが挙げられる．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．

しかし，これまでのマルチエージェント環境における学習を取り扱った研究には，他者の行動に関する仮定が理想的なものが多かった．Littman [3] は，格子環境下においてマルコフゲームの枠組みを応用した強化学習を提案している．ここで，学習者の評価関数は 2 人ゼロ和の関係にあり，学習者は常に他者の最悪の行動を想定することになり，協調の問題には適用できない．また，他者の行動が直接観測される必要がある．Sandholm and Crites [4] は 繰り返しの囚人のジレンマ問題に強化学習を適用し，学習が成功するためには，十分な過去の観測量と行動が必要であることを示した．しかし，その履歴の長さを決定するのは，一般に困難な問題である．

また，マルチエージェントを研究する題材として近年ロボカップ [1] があるが，その中で学習の問題を取り扱った研究として Stone and Veloso [5] や，Uchibe et al. [6] がある．Stone and Veloso は layerd learning という，階層構造の学習法を提案し，Uchibe et al. は複数行動の調停問題を解決する学習法を提案しているが，瞬間のセンサ情報 (観測量) 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，複数ロボットの学習問題に適用することは困難である．

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，その結果をもとに強化学習をおこなう手法を提案する．また，マルチエージェント系での学習を安定にするための学習のスケジューリング法を提案する．

提案する手法を簡単な 1 対 1 のサッカーゲームに適用する．環境には 2 台のロボットが存在し，それぞれに対して異なるタスクを与える．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．各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　学習アルゴリズム

2.1　アーキテクチャ

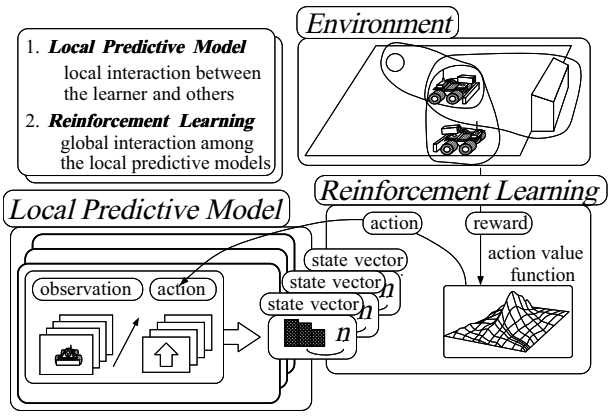


図 1: Proposed architecture

図 1 は各ロボットに与えられる行動獲得のためのアーキテクチャである．はじめに，学習者はセンサ情報だけでなく，学習者自身の行動のシーケンスから局所予測モデルを構築する．局所予測モデルは対象の次の運動が予測できるような状態ベクトルを推定する．次に推定された状態ベクトルをもとに，協調行動を獲得のための学習を開始する．

2.2 局所予測モデル

局所予測モデルは，多入力 (行動) 多出力 (観測) の関係を記述する必要がある．状態表現の方法として，システム同定の一つである正準変量解析 (Canonical Variate Analysis) [2] を用いて，局所予測モデルを構築する．CVA を用いた局所予測モデルの詳細については文献 [7] を参照されるとして，ここでは簡単な概略だけを述べる．

CVA は離散時間で線形の状態空間モデル

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

を用いる．ここで $\boldsymbol{u}(t) \in \Re^m$ と $\boldsymbol{y}(t) \in \Re^q$ はそれぞれロボットの行動ベクトルと観測ベクトルであり，$\boldsymbol{x}(t) \in \Re^n$ は状態ベクトルである．また，$\boldsymbol{A} \in \Re^{n\times n}$, $\boldsymbol{B} \in \Re^{n\times m}$, $\boldsymbol{C} \in \Re^{q\times n}$, $\boldsymbol{D} \in \Re^{q\times m}$ はパラメータ行列である．学習者は観測と行動のシーケンス $\{\boldsymbol{y}, \boldsymbol{u}\}$ から状態ベクトルを次数を含めて推定しなければならない．状態ベクトル $\boldsymbol{x}$ は過去の観測と行動のシーケンスの線形和

$$\boldsymbol{x}(t) = [\boldsymbol{I}_n \ \boldsymbol{0}]\boldsymbol{U}\boldsymbol{p}(t), \tag{2}$$

によって状態を表現する．ここで，

$$\boldsymbol{p}(t) = [\boldsymbol{u}(t-1) \cdots \boldsymbol{u}(t-l)\, \boldsymbol{y}(t-1) \cdots \boldsymbol{y}(t-l)]^T,$$

であり，$\boldsymbol{U} \in \Re^{l(m+q)\times l(m+q)}$ はCVA によって計算される行列であり，l はこう慮する履歴長さである．また n は状態ベクトルの次数であり，情報量規準によって決定する．

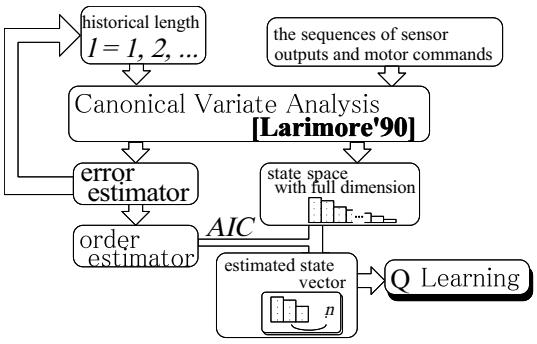


図 2: Local predictive model

図 2 に局所予測モデルを示す．

2.3 モジュール強化学習

局所予測モデルは，学習者と他者との局所的な相互作用を推定するだけであるため，ロボットは複数の局所予測モデルと与えられたタスクとの大局的な相互作用を推定する必要がある．大局的な相互作用を学習するために，モジュール強化学習 [6] を用いる．モジュール強化学習は，n 個の学習結果を統合し，適切な行動を獲得する学習アルゴリズムである．

モジュール強化学習の基本的な考え方を示す．ただし，図 3 は $n = 2$ の場合を表している．モジュー

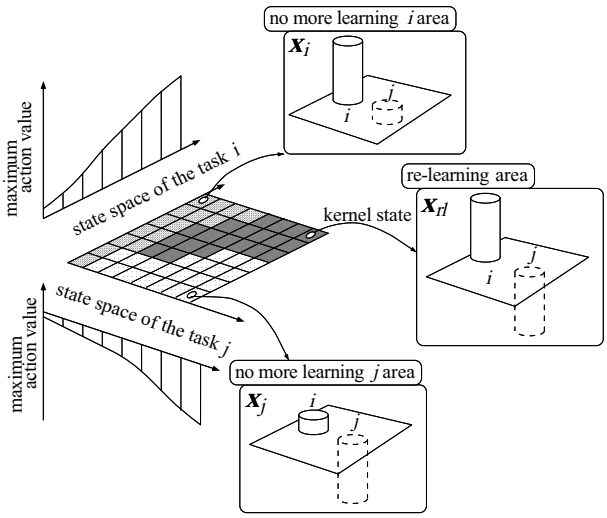


図 3: Basic idea of the modular reinforcement learning

ル強化学習では，学習時間を削減するため，全状態空間を

- no more learning area
  以前に学習した結果をそのまま適用できる部分状態 → n 個
- re-learning area
  学習結果が干渉しており，再学習を必要とする部分状態 → 1 個

の部分状態に分類する．分類は最大の行動価値関数をもとにクラスタリングされる．

no more learning area に属する状態では，すでに獲得された行動価値関数をそのまま利用する．re-learning area に属する状態には，通常の Q 学習を

適用するが，Q 学習の減衰パラメータ γ を適切に設定する必要がある．モジュール強化学習では，γ を no more learning area の Q 値を境界条件となるように推定しながら，re-learning area の行動価値関数を再学習する．結果として，モジュール強化学習は複数の行動の調停を，学習時間とパフォーマンスのトレードオフを考慮した行動が獲得できる．

3 複数ロボットのための学習のスケジュール

複数ロボットが存在する環境下で協調行動を獲得させるために，学習のスケジューリングをおこなった．一般に，ロボットの学習は次の 3 通りに分類することができる．

1. 実環境だけで学習：
   単純な環境で単純なタスクである場合を除いて，現実的ではない．
2. 計算機上の学習結果を実ロボットに適用：
   計算機上でのシミュレーションと実環境にはギャップがあり，いくらかの修正を必要とする．
3. 計算機上で獲得された結果を実環境で修正：
   シミュレーション結果をもとにして，実環境での学習をスケジューリングする．

ここでは，3 番目の方法を採用する．

図 4 に学習のスケジュールを示す．最初に局所予測モデルの構築を計算機上でおこなう．局所予測モデルは，各ロボットが同時に推定し，その時の行動戦略はランダムである．次に，推定結果をもとに学習を開始するが，学習の初期段階でのランダム性を排除するため，学習するロボットを 1 台指定し，それ以外のロボットの行動戦略を固定する．学習ロボットの学習が終了した後で，別のロボットの学習を開始する．計算機シミュレーション上で局所予測モデルの更新と行動価値関数の更新を繰り返すことで，各ロボットは目的の行動を獲得する．

次に，学習結果を実ロボットに適用し，そのときの結果をもとに局所予測モデルを更新する．シミュレーションの結果を初期値とすることで，実環境での探索を短縮できる．局所予測モデルを構築した後で，シミュレーションと同様にして行動学習をおこなう．

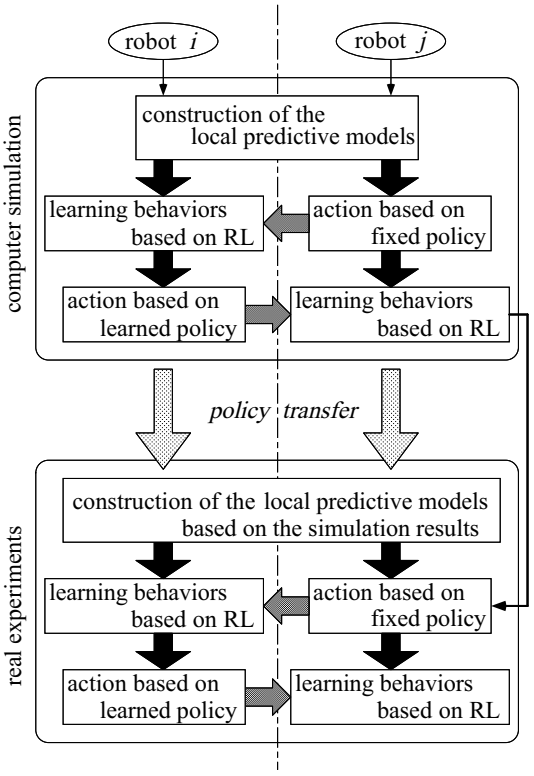


図 4: Schedule for learning in multi robots environments

4 タスク

提案手法を，2 台のロボットが存在する環境下での，簡単なサッカーゲームに適用した (図5)．各ロボットは TV カメラを一つ搭載し，そこから得られる画像情報から環境の状況を観測する．

モータコマンドとして，各ロボットは 2 自由度を持つ．そこで，ロボットへの制御入力 $\boldsymbol{u}$ は 2 次元ベクトル

$$\boldsymbol{u}^T = [v \ \ \phi], \quad v, \phi \in \{-1, 0, 1\},$$

として表現する．ここで，v は台車の移動速度であり，ϕ はステアリングの角度である．また，各ロボットが観測できる画像特徴量 (観測ベクトル) を図6 に示す．結果として，ボール，ゴール，ロボットに関する観測ベクトルの次数はそれぞれ 4, 11, 5 となる．詳細は [7] を参照されたい．

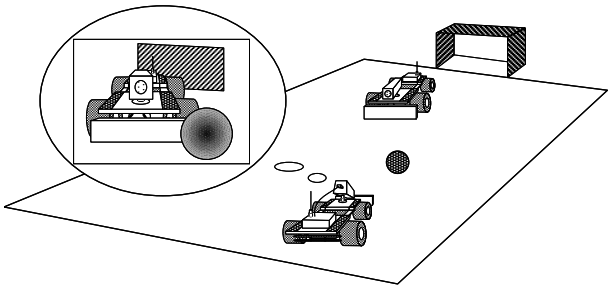
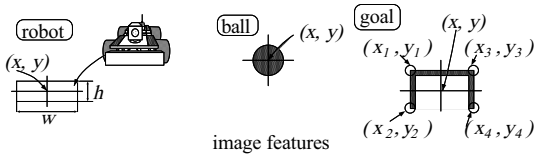


図 5: robots and environment



robot	ball	goal
area	area	area
center position	center position	center position
height	radius	4 corners
width		

図 6: Image features of the ball, goal, and agent

5 実験結果

最初にシューターとパッサーは，ボール，ゴール，そして互いの局所予測モデルを，計算機のシミュレーション上で構築する．次に，シューターを静止させた状況下で，パッサーは行動の学習を開始する．パッサーの学習が終了した時点で，パッサーの行動政策を固定し，シューターの学習を開始する．パッサーは，ボールをシューターにパスしたときに報酬 1 を受け取り，シューターはボールをゴールにシュートしたときに，報酬 1 を受け取る．さらに，ロボット間で衝突が発生した場合，−0.3 の報酬が与えられる．

計算機上での学習が終了した時点で，獲得された結果を実ロボットに適用する．実環境での局所予測モデルを再構築する為の行動戦略は，80 % の確率でシミュレーションで獲得された行動戦略を用い，20 % の確率で，ランダムに行動する．局所予測モデルを構築するために，実環境で 100 回の試行をおこなった．局所予測モデルが更新された後で，ロボットは行動価値関数を収集した実データをもとに洗練する．最後に，実環境でのパフォーマンスを計測するため，

表 1: The estimated dimension

observer	target	l	n	$\log \lvert \boldsymbol{R} \rvert$	AIC
computer simulation					
shooter	ball	2	4	0.23	138
	goal	1	2	−0.01	121
	passer	3	6	1.22	210
passer	ball	2	4	0.78	142
	shooter	3	5	0.85	198
real experiments					
shooter	ball	4	4	1.88	284
	goal	1	3	−1.73	−817
	passer	5	4	3.43	329
passer	ball	4	4	1.36	173
	shooter	5	4	2.17	284

表 2: Performance result in real experiments

	before learning	after learning
success of shooting	57/100	32/50
success of passing	30/100	22/50
number of collisions	25/100	6/50
average steps	563	483

50 回の試行をおこなった．

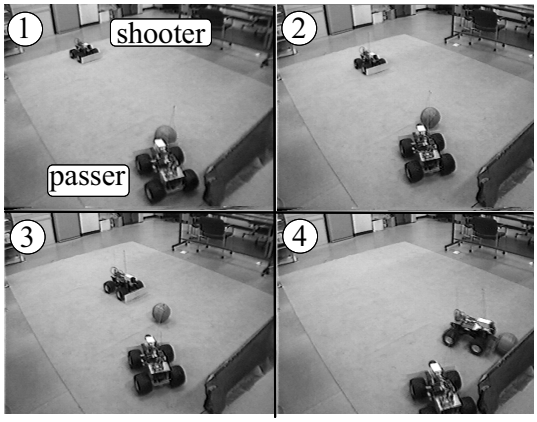
表 1 は計算機シミュレーションおよび実環境での推定された状態ベクトルである．ここで，$\boldsymbol{R}$ は局所予測モデルの誤差の共分散行列であり，AIC は赤池の情報量である．例えばゴールの場合，次の状況を予測するための履歴長さ l は $l = 1$ で充分であるのに対し，ボールの場合は $l = 2$ ステップの履歴長さが必要である．計算機と実環境で推定された状態ベクトルの次数が異なる理由として，

- ノイズのために，実環境での予測誤差は，計算機上での予測誤差よりもずっと大きく，状態の次数を増加しても推定精度が向上されない．
- 局所予測モデルを構築するための観測と行動のシーケンスが異なっている．計算機上ではランダムに行動することが可能であるが，実環境では計算機上での学習結果に基づき行動するため，得られるデータにタスク依存の偏りがある．

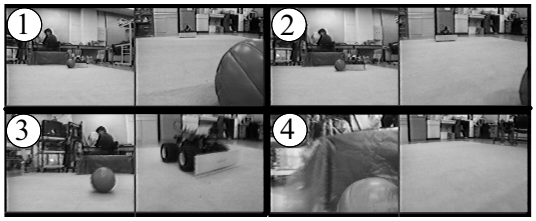
ことが挙げられる．結果として，実環境での履歴長さ l は，シミュレーションでの履歴長さよりも長くなっている．一方で，状態ベクトルの推定次数 n はシミュレーションよりも実環境の方が小さいことがわかる．

次に獲得されたビヘービアのパフォーマンスについて述べる．センサ情報だけを状態量として用いた場合との比較は [7] を参照されるとして，ここでは，実環境でのパフォーマンスの洗練の結果について述べる．表 2 に，学習前後のパフォーマンスの比較結果を示す．これからシミュレーション結果をそのまま適用するよりも，パフォーマンスが改善されていることがわかる．特に，衝突回数と目標達成までの平均ステップ数が改善されている．

また，実環境で獲得されたシューターとパッサーの局所予測モデルを交換した場合のビヘービアを観察したところ，両方のロボットにおいて局所予測モデルの予測誤差が大きくなり，適切な行動を生成できなかった．このことは，実ロボットにおいては推定された局所予測モデルは交換不可能なことを示している．最後に，図 7 に獲得されたビヘービアの例を示す．まず，パッサーがボールをシューターに向かってボールを蹴り，シューターはボールをゴールにシュートする．パッサーはボールを蹴った後は，シューターとの衝突を回避するための行動をしていることがわかる．



(a) top view



(b) obtained image

図 7: Acquired behavior

6 おわりに

本稿では，強化学習を複数のロボットが存在する環境下に適用するための手法について提案した．今後の方針として，提案手法を拡張し，3 台以上のロボットが存在する環境下で，協調，競合行動を学習させるタスクが考えられる．

参考文献

[1] H. Kitano, M. Asada, Y. Kuniyoshi, I. Noda, E. Osawa, and H. Matsubara. Robocup a challenge problem for ai. *AI Magazine*, 18(1):73–85, 1997.

[2] 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.

[3] 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.

[4] T. W. Sandholm and R. H. Crites. On multiagent Q-learning in a semi-competitive domain. In *Workshop Notes of Adaptation and Learning in Multiagent Systems Workshop, IJCAI-95*, 1995.

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

[6] 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.

[7] E. Uchibe, M. Asada, and K. Hosoda. State space construction for behavior acquisition in multi agent environments with vision and action. In *Proc. of International Conference on Computer Vision*, pp. 870–875, 1998.