3W－5

確率的傾斜法を用いた状況領域における適応学習

重永 稔朗　中西正和
慶應義塾大学大学院 理工学研究科 計算機科学専攻

1. はじめに

近年，分散人工知能研究の一つとして，マルチエージェント環境に対し，強化学習を用いることによって適応的に学習する方法が行われている．

一般的にマルチエージェント環境は，状態表現の複雑さ，状態遷移とイベントの不確実さ，学習試行数，強化とフィードバック，複数目的という問題がある．この動的で複雑な環境を状況領域(situated domain)においての学習という．伝統的な強化学習の枠組では状況領域に必ずしも対応できないと言われている．

本研究では，①単順に「観測入力に対する行動出力の確率」の関数を最適化する確率的傾斜法[1]という強化学習法を用い，②状況領域を考慮した，マルチエージェント環境における適応学習を試みる．

2. 状況領域の学習

図1のマルチエージェント環境を考える．

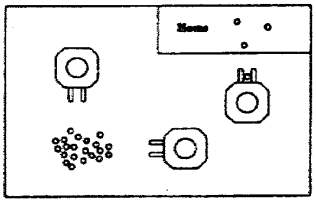

図 1: ゴミ集めのタスク

2.1 学習タスク

学習タスクは集団行動のためのもっとも効果的な条件と行動器のマッピングを見つけることから成る．それぞれのエージェントはpuckを見つけてhomeへ持っていくために，それぞれの条件のための最適行動を選択することを学習する．エージェントの行動は，basic behavior approach(基本行動アプローチ)[2, 3]に基づいて設計されており，マルチエージェントにおいて基本的な行動群であることが示されている．レパートリは次の行動から成る．

safe-wandering　dispersion　homing

行動を引き起こすための必要十分な条件の空間のみ考慮することによって，状態空間は次の状態変数のcross-productに減少される(図2)．

at-home?　have-puck?　near-intruder?

puckを掴む，または落とすための条件は組み込んである．エージェントが指の間にpuckを検知した場合それを掴む．類似して，エージェントがhomeに達した場合，puckを持っていれば落とす．

	condition(C)		
	at-home?	have-puck?	near-intruder?
X_1	0	0	0
X_2	0	0	1
X_3	0	1	0
⋮			

behavior(B)	
b_1	wandering
b_2	dispersion
b_3	homing

図 2: 条件-行動対 $A(X_i, b_j)$

3. 確率的傾斜法

エージェントの学習目標は，報酬獲得の平均を最大化するように，それぞれの観測において行動を選択する確率分布を形成することである．観測 X においてエージェントが行動 a を選択する確率を政策 π と呼び，関数 $\pi(a_t, W, X_t)$ で表す．パラメータ W はエージェントの内部変数ベクトルを表す．エージェントは内部変数 W を調節し，確率的政策 π を変えることで学習が進む．

3.1 エージェントの構造と学習アルゴリズム

1. 環境の観測 X_t を受け取る．
2. $\pi(a_t, W, X_t)$ の確率で行動 a_t を実行する．
3. 環境から報酬 r_t を受け取る．
4. 内部変数 W の全ての要素 w_{ij} について以下の適正度 $e_{ij}(t)$ と適正度の履歴 $D_{ij}(t)$ を求める．ただし，γ は割引率 $(0 \leq \gamma < 1)$ である．

$$e_{ij}(t) = \frac{\partial}{\partial w_{ij}} \ln(\pi(a_t, W, X_t))$$
$$D_{ij}(t) = e_{ij}(t) + \gamma D_{ij}(t-1)$$

Adaptive Learning in Situated Domains with Stochastic Gradient Ascent
Toshiro SHIGENAGA　Masakazu NAKANISHI
Department of Computer Science, Faculty of Science and Technology, Keio University 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223, Japan

表 1: 学習結果の例

	condition			agent1			agent2			agent3		
	HO	HP	NI	wander	disperse	homing	wander	disperse	homing	wander	disperse	homing
X_1	0	0	0	49.99	49.99	0.02	52.86	0.00	47.14	99.95	0.03	0.02
X_2	0	0	1	33.71	38.58	27.71	33.10	33.01	33.89	58.82	24.97	16.22
X_3	0	1	0	33.11	0.03	66.86	49.17	0.01	50.82	33.09	0.06	66.85
X_4	0	1	1	99.86	0.07	0.07	33.28	33.04	33.69	33.35	66.59	0.06
X_5	1	0	0	99.99	0.00	0.01	31.08	34.29	34.63	20.67	43.44	35.89
X_6	1	0	1	28.15	38.25	33.61	33.23	33.57	33.20	45.58	30.94	23.48

5. 以下の式を用いて $\Delta w_{ij}(t)$ を求める.

$$\Delta w_{ij}(t) = (r_t - b)D_{ij}(t)$$

6. 政策の改善: 以下の式で全ての w_{ij} を更新する.

$$w_{ij} \leftarrow w_{ij} + \alpha(1-\gamma)\Delta w_{ij}(t)$$

7. 時間ステップ t を $t+1$ へ進め, 1へ戻る.

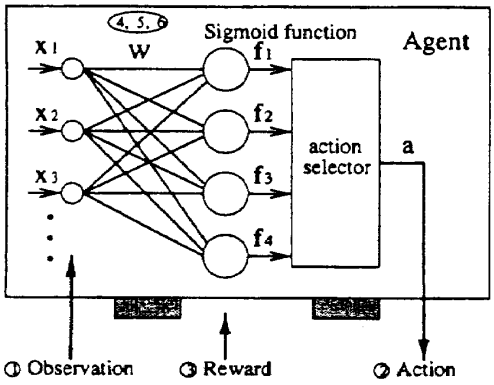

図 3: エージェントの内部構造

報酬は puck を掴んだとき, puck を持っている場合に home に近付いたとき得られる. また, 各パラメータは $\gamma = 0.9, b = 0.01, \alpha = 0.3$ とした.

図3のようなエージェントの内部構造と, 1から7の処理を繰り返すことによって, 報酬獲得に関係ない行動は打ち消され, 報酬獲得に関係する行動だけが強化される. 行動の履歴を強化するので, 報酬の獲得に遅れのある行動も強化される.

4. 実験および考察

エージェントの学習特性を調べるために, ゴミ集めのタスクを取り上げ, そのシミュレータ **MARS** [4] を通して実験を行った. これは, 現実を単純化しているものの, 実機との対応付けは十分配慮してある.

エージェント数が3のときの個々のエージェントの学習結果を表1に示す. 表中の数値は各行動の選択確率を百分率で表したものである. 環境の puck 数は7で各エージェントのステップ数は 3000 回とした.

何も観測されない場合は `safe-wandering`, puck を掴んだ場合は `homing` という行動が強化されている. これはゴミ集めのタスクの基本的な行動を学習していることが分かる. 状態 X_2 のときは `dispersion` が学習されると考えられるが, 実際はランダムな行動をとっている. 状態 X_4 の場合, エージェント1は `safe-wandering`, エージェント3は `dispersion` を学習し, 一方が進み他方が回避する協調行動が見られた. また, エージェント2は puck を見つけるのが遅く, 全般的に学習が進まなかった. そのため, この環境内では冗長エージェントとなった.

10回の試行における puck の平均獲得数は 5.1 個であった. 他の強化学習法として Q-learning を用いた場合 3.1 個であった. 確率的傾斜法の頑健性が寄与していると考えられる.

5. おわりに

本稿では確率的傾斜法と basic behavior approach を用いたエージェントアーキテクチャを提案し, 状況領域の適応学習について述べた. 今後の課題としてヘテロエージェント系の試みが挙げられる.

参考文献

[1] Kimura, H., Yamamura, M. and Kobayashi, S.: Reinforcement Learning in POMDPs with Function Approximation, *Proceedings of the 14th International Conference on Machine Learning* (1997).

[2] Matarić, M.: Designing Emergent Behaviors: From Local Interactions to Collective Intelligence, *From Animals to Animats: International Conference on Simulation of Adaptive Behavior* (1992).

[3] Matarić, M.: Reinforcement Learning in the Multi-Robot Domain, *Autonomous Robots*, Vol. 4, No. 1, pp. 73–83 (1997).

[4] 松原仁, 開一夫, 木村陽一, 國吉康夫: マルチ・エージェントシステムにおける学習への実験的アプローチ, マルチエージェントと協調計算 IV, pp. 177–184, 日本ソフトウェア科学会 MACC'94, 近代科学社 (1994).