自己組織化ファジィニューラルネットワーク
を用いた強化学習システムに関する研究

Study on Reinforcement Learning System using Self-Organizing Fuzzy Neural Network

平　成 26 年 3 月

呉　本　　尭

山口大学大学院理工学研究科

学位論文の要旨

人工知能分野に属する機械学習の一つである強化学習に関して、1940 年代に Bellman が提案した動的計画法に基づいたいくつかの学習アルゴリズム、即ち、Actor-critic，Q，Sarsa 等の学習法が提案されている。近年では、これらの学習法を用いた強化学習の応用も盛んに行われてきており、人間を相手にするサービス型ロボット等、従来型の制御では困難な、環境変化に柔軟に対応可能な応用研究も数多くなされてきている。しかしながら、例えば、自然界で発生する現象は時間的・空間的に連続な事象であるのに対し、Q，Sarsa 等の学習法は離散的な現象を対象としている。本論文では、これら、それぞれの学習法に対して、連続的な現象を取り扱えるニューロファジィ型強化学習システムを提案している。そして、これまでは、単独のエージェント（ロボット）を学習の対象とした研究が殆どで有るのに対し、本論文ではエージェント群の効率的な群学習アルゴリズムを提案し、これが単独学習よりも有用である分野を例示している。

本論文の構成は、以下の通りである。

第 1 章では、研究の背景と位置付け、及び論文の構成について述べる。

第 2 章では、知的個体の概念と機械学習の概要を述べ、本論文の研究対象分野であるファジィ推論、ニューラルネットワーク及び強化学習を概説する。特に、観測情報を分類するための自己組織化型ファジィニューラルネットワーク（SOFNN）を提案する。ここで、「自己組織化」とは、入力データ駆動によるネットワークの自動形成を意味する。多次元入力空間による入力に対し、提案手法では、閾値制御及びルール生成規則によって、ファジィメンバーシップ関数や、ファジィルールを生成・結合し、ファジィニューラルネットワーク（FNN）を構成する。

第 3 章では、FNN を用いた学習アルゴリズムの異なる三種類の強化学習システムを提案している。

(i) FNN を用いた Actor-Critic 型強化学習システム（FAC）

SOFNN の出力に荷重を加え、ネットワークの一層とする状態価値関数 Critic と行動価値関数 Actor を可塑的に結合する。行動価値関数の出力は、確率探索をもたらす行動選択方策関数に用いられ、知的個体が行動を選択し、出力する。行動の結果による環境の状態の変化と、知的個体に返す報酬や罰を含む時間差分誤差（TD-error）を用いて、SOFNN と状態価値関数・行動

価値関数の結合荷重を修正することよって、知的個体が価値の高い状態へ遷移するように、行動方策が修正される。座標情報（離散値及び連続値）を用いた有限マルコフ決定過程(MDP)を持つ目標探索問題のシミュレーションを行い、提案した FAC によって構成された知的個体が環境との相互作用の結果より、適切な行動を獲得できることが認められた。

(ii) FNN を用いた Q 学習型強化学習システム（FQ）

システムの構成は FAC と異なり、SOFNN の出力は、状態―行動価値関数（Q 関数）と可塑的に結合する。行動選択方策関数は Q 関数を用いて構成される。1989 年に Watkins により提案された一般的に利用されている Q 学習アルゴリズムと異なり、従来の「TD 誤差を直接に Q 値の修正に用いる」の代わりに、TD 誤差を FAC の結合荷重の修正に導入し、システム（FQ）の出力改善につなげる。近傍情報しか得られない部分観測マルコフ決定過程(POMDP)を持つ目標探索問題のシミュレーションを通して、提案した FQ によって構成された知的個体が環境との相互作用の結果より、適切な行動を獲得できることが認められた。

(iii) FNN を用いた Sarsa 型強化学習システム（FS）

FS の構成は FQ と同じであるが、Q 学習の TD 誤差の計算法と異なり、1994 年に Rummery & Nirajan により提案された Sarsa 学習の TD 誤差式を FS の結合荷重の修正に導入する。また、POMDP 下の目標探索問題のシミュレーションを通して、提案した FS によって構成された知的個体が環境との相互作用の結果より、適切な行動を獲得できることが認められた。

また、魚や鳥など生物の群行動のシミュレーションができる行動選択ルールを未知環境における目標探索問題の解法に導入し、知的個体間の適切な距離を保つように、「離れず近すぎず」行動の報酬を個体の価値関数に反映し、「群学習」によって、最適解、または準最適解をより早く発見することを図る。「群学習」の概念をそれぞれの提案強化学習システムに導入した場合と、他の個体との距離をを考慮しない「単独学習」の場合の比較が、シミュレーションの結果によって行われた。

第 4 章では、不完全観測環境（POMDP）における目標探索問題のシミュレーションを用いて、ランダム探索、従来の Q 学習法、従来の Sarsa 法、及び提案した FAC、FQ と FS のそれぞれの学習結果・学習性能の比較を行い、提案法の有効性を確認する。また、提案法におけるパラメータ設定方法について考察する。

最後に、第 5 章では、本論文のまとめと今後の課題について述べる。

Abstract

Individuals with low intelligence ability such as insects, fishes, and birds can collect together as swarms, schools and flocks and show complex behavior patterns. Imitating these behaviors, meta-heuristics methods, so-called “swarm intelligence”, have been proposed in last decades to find the optimal solution of mathematical problems. For example, “particle swarm optimization (PSO)” (Kennedy & Eberhart 1995) and “ant colony optimization (ACO)” (Dorigo 1999) are well known. The former is useful to find the approximate solution of nonlinear functions and the later deals with the optimal combination problems such as “travelling salesman problem (TSP)”. “multi-agent systems (MASs)” are also defined as distribution processing models to explore unknown environments, to diagnose large scale systems, to simulate social problems, to realize artificial lives and so on. Agents, i.e. individuals in MASs, are autonomous entities with abilities of state recognition, decision making, active learning, etc.

However, individuals designed in the conventional swarm intelligence are generally with simple structures and low intelligent functions, lacks of abilities of higher animals such as unknown information classification, adaptive behavior acquisition.

Meanwhile, artificial neural networks (ANNs), fuzzy inference systems (FISs) and the fusion systems of them called neuro-fuzzy systems have been widely used in the adaptive/intelligent control of autonomous robots. Additionally, reinforcement learning (RL) provides kinds of powerful machine learning algorithms which make robots more autonomously.

However, because multiple autonomous mobile robots act to affect each other, the transition of states become to non-Markov decision processes, so the conventional RL is hard to deal with them (MASs). Recent approaches usually use graph topology method (Jababaie et. al 2003) or nonlinear dynamics models (Moreau 2005) hiring complex mathematical formulae sensitive to disturbance.

In this paper, a novel reinforcement learning system with self-organizing neuro-fuzzy network is proposed. As an internal model of autonomous agents, the proposed system aims to tackle the problems exist in conventional swarm intelligence and swarm control of autonomous robots. “self-organizing fuzzy neural network (SOFNN)” is a fuzzy neural network (FNN) which structure, such as membership functions, fuzzy rules, and the connections between them, is generated by the input data automatically. The output of SOFNN are connected to a state value function V, an action value function A or a state-action value function Q. Agents classify the input pattern as a certain state by FN and decide an adaptive action using a stochastic policy effected by value functions A or Q. The learning to acquire adaptive behavior is realized by adopting 3 kinds of classic reinforcement learning (RL) algorithms: Actor-Critic learning (Barto

et. al 1983; Sutton 1988), Q learning (Watkins 1989) and Sarsa learning (Rummery & Niranjan 1994; Sutton 1996). Using a rule of birds flocks BOID, i.e., individuals keep suitable distance between each other, they are able to explore the unknown environment, to find the appropriate solution more efficiently, and to acquire adaptive behaviors after training process during their exploration.

In detail, the 3 kinds of reinforcement learning systems with SOFNN proposed in the paper are as follows:

Model 1: An Actor-Critic type reinforcement learning system with neuro-fuzzy network (called "FAC");

Model 2: A Q-Learning type reinforcement learning system with neuro-fuzzy networks (called "FQ");

Model 3: A Sarsa-Learning type reinforcement learning system with neuro-fuzzy networks (called "FS").

Model 1 (FAC) is an online processing system which uses a state-value function (Critic) and a action-value function (Actor) both connected to the output of Fuzzy Net with adjustable weights. "Temporal difference error" (TD error) is used to learning algorithm which modifies the weights of connections and probability distribution of action selection (i.e., "policy function" called in RL). The input space (states) and output space (actions) of the system may be discrete or continuous. Furthermore, agents (intelligent individuals) defined by FAC are able to explore unknown environments, not only in the case of observable Markov decision process (MDP) such as the input are global coordinates information, but also partial Markov decision process (POMDP) such as local observable environment.

Model 2 (FQ) uses Q-Learning algorithm which stresses the exploitation of learning history. It is a policy-off RL system and the advantage of this model is its fast approach to the optimal solution, whereas the learning convergence is affected by the input of states in POMDP environment.

Model 3 (FS) has the same structure with the FQ but uses Sarsa Learning algorithm instead of Q-learning algorithm. FS is a policy-on RL system, current state and the next state are observed and it makes the learner (agent) behave "more careful" exploration comparing with FQ.

When the rewards which are evaluation to the qualification of distance between multiple agents, "swarm learning", which means the situation of suitable distance between agents is encouraged with positive rewards, yields the collective behaviors of agents and higher learning performance comparing with "individual learning", the opposite case. Simulations of goal-directed exploration problems showed the effectiveness of the proposed systems. And the contribution of this paper may be applied to multiple autonomous robots control which is useful in the fields of space / deep sea exploration, and other tasks in the extreme environments.

目　次

表目次

図目次

概　要

本論文では、知的個体の内部モデルとして、ファジィニューラルネットワークを用いた強化学習システムを提案する。また、提案した内部モデルを持つ知的個体が複数存在する場合に、群れの形成および適応群行動の獲得について論述する。

ここで、知的個体とは、「未知環境に適応できる行動主体」であり、適応能力を持たない一般的な自律エージェントに比べ、より環境の観測・認知及び行動の学習能力が高い「インテリジェントシステム」を指す。また、内部モデルとは、知的個体の状態認識、方策構成および行動出力などの機能を実現する数理モデルであり、自律ロボットなど知的人工物への応用が可能である。

一般的に、機械学習は、人工知能分野に属し、観測・計測によって得られたサンプルデータに対し、推論や帰納などの処理によって、それらのデータの規則や、知識を獲得する過程である。その計算的な処理の方式によって、機械学習は概ね「教師あり学習」、「教師なし学習」及び「強化学習」に分類できる。中でも、認知対象である「環境」に関する事前情報を必要とせず、学習主体が環境との相互作用と、その過程に伴う報酬や罰を通して、最適な「行動方策」を学習によって獲得する強化学習方式が、近年、知的制御や自律ロボットの開発などの分野で脚光を浴びつつある。

これまでの強化学習は、主に三つのアプローチがある。(i)入出力の対応関係を発見しようとする分類子システムや経験強化型 Profit Sharing ; (ii) 関数近似器やニューラルネットワークなどを用いた環境同定（モデルベース）によるモデル学習 ; (iii)動的計画法に基づく方策修正学習である。これらのアプローチは、いずれも学習主体の行動結果から、様々な入力パターンに対応する適切な出力を求めようとする強化学習の基本的な原理に基づくものである。一方、実ロボットや複数のエージェントが存在するマルチエージェントシステムでは、入力パターンの不完全性、不確定性および膨大性、いわゆる「(状態の) 次元の呪い」から、従来の強化学習手法でそのまま対応することは困難である。これらの意欲的な課題に対し、近年、様々な強化学習システムが提案されているが、問題の複雑さから、実用性や解の収束性、ロバスト性などの問題は依然解決されていない。

本論文では、まず、観測情報を分類するため、自己組織化ファジィニューラルネットワーク(SOFNN: self-organizing fuzzy neural network)を構築する。ここで、「自己組織化」とは入力データ駆動によるネットワークの自動形成を意味する。多次元入力空間による入力に対し、提案手法では、閾値制御及びルール生成規則によって、ファジィメンバーシップ関数や、ファジィルールを生成・結合し、自動的にファジィニューラルネットワークを構成

する。

次に、SOFNN の出力に荷重を加え、状態価値関数、行動価値関数、または状態―行動価値関数と可塑的に結合する。これらの価値関数を用いて強化学習の確率的な行動方策を構成する。Actor-Critic 学習や Sarsa 学習、Q 学習といった従来の強化学習アルゴリズムを FNN と価値関数間の結合荷重の修正に導入し、行動方策の修正を行う。よって、分類された環境情報、すなわち、観測状態に応じる適切な行動を選択することが可能となる。なお、本論文において、学習方式の異なる強化学習システムを構成しており、それぞれ、「FNN を用いた Actor-Critic 型強化学習システム(FAC)」、「FNN を用いた Q 学習型強化学習システム(FQ)」及び「FNN を用いた Sarsa 学習型強化学習システム(FS)」と称す。

また、構築された各強化学習システムを知的個体の内部モデルとして、知的個体群によって、目標探索問題をいかに効率よく解決するかを考案する。魚や鳥などの群行動の原理を未知環境における目標探索問題の解法に導入し、「個体間の距離が離れず近すぎず」という行動効果を知的個体の価値関数に反映し、「群学習」によって、最適解、または準最適解をより早く発見することを図る。

未知環境における目標探索問題の計算機シミュレーションにおいて、観測する状態空間の表現に応じて、提案する各学習システムを使い分ける。絶対位置の分かる完全観測空間の場合は、座標情報（離散値及び連続値）が入力され、Actor-Critic 型システムである FAC が対応する。近傍情報しか得られない部分観測空間の場合は、近傍観測入力（離散値）に対して、Q 学習型システム FQ または Sarsa 学習型のシステム FS が対応できる。

最後に、考察と結論を通して、本論文の成果及び今後の課題について述べる。

第 1 章

序　論

1.1　研究の背景と位置付け

電子計算機から知能を持つ人工物へ

1940 年代に電子計算機が誕生してから、入力データに対する蓄積（記憶）、計算、推論などの情報処理が可能となった。データ処理の自動化を超え、人間の知能に匹敵するような機械をも構築しようとする研究は 1956 年から「**人工知能**（AI: Artificial Intelligence）」と呼ばれるようになり、半世紀以上の歴史がある。

古典的な人工知能は、事象の記号的な記述、論理的な帰納、統計的な推論などのアプローチが多いが、近年、計算機の性能の飛躍的な向上により、**計算知能**(CI: Computational Intelligence)が盛んとなり、人工神経回路網（ニューラルネットワーク）や進化的計算を始め、データマイニングや群知能などの新たな人工知能分野が生まれた。近い将来、「人間なのか、マシンなのか」というチューリングテスト(Turing Test)に合格するインテリジェントシステムが、いずれは完成されるであろう。

1997 年 5 月、IBM のチェスマシン Deep Blue がチェスの人間世界チャンピオン Kasparov と対戦し、勝利した[1]。Deep Blue は 32 体（VLSI プロセッサ計 512 個）のスーパーコンピュータを並列に使用し、1 秒間に数億のチェスの局面を読むことができ、10 手先以上優位になる手を決める能力を持っていた。しかし、そのチェスマシンを開発したチームは人工知能技術を使用していないという。

2013 年 4 月、日本の将棋のプロ棋士 5 人が、5 種類のコンピュータソフトと対戦した団体戦の末、1 勝 3 敗 1 引き分けで敗れた。トップレベルの棋士三浦弘行八段が負けた東京大学の教員と学生が開発した「GPS 将棋」は、690 体のコンピュータと結んで、1 秒間に 2 億 5 千万手を読み、記録した過去の棋譜を引き出して使用するだけでなく、学習もできるという。チェスとルールが違い、相手から奪った駒が使える将棋の盤面の数は 10^{220} にも上り、これまで人間の勝算が大きいと思われていただけに、早くも人工知能の勝利が現実味を帯びてきた。

1997 年、人間を超える人工知能研究のもう一つの目標「サッカーロボットプロジェ

クト」が提案された。それは RoboCup プロジェクトと呼ばれ、自律移動サッカーロボットが、2050 年に「人間のワールドカップ優勝チームと対戦し、勝利する」という目標を掲げ、国際会議と国際競技大会が毎年開催されている[2]。

サッカープレイヤーとしてのロボットは、身体能力は別として、自らが置かれている状況を知覚・判別することと、その先を予測し、個体としての自己やチーム全体のために適切な行動を決定することが要求される。このような物理的なロボット及びその「脳」に相当するソフトウェアは、抽象的に「**自律エージェント**(Autonomous Agent)」、或いは、「**エージェント**（Agent）」と呼ばれ、人工知能分野の新たな概念として、1990 年代から急速に注目されてきた[3]。自律エージェントは「未知環境に適応するような行動を自ら見つける」ことができる知能を持つ**人工物**として、これまで、幅広い分野へ応用されている。例えば、電子秘書、スケジューリング、情報検索、通信技術、大規模システム診断、自律ロボットの設計などが挙げられる[3]-[13]。

自律エージェントの学習

自律性と知的能力の獲得または向上のため、自律エージェントには学習機能が要求される。これまで、人工知能分野の「**機械学習**(Machine Learning)」と呼ばれる学習方式は、概ね「**教師あり学習**(Supervised Learning)」、「**教師なし学習**(Unsupervised Learning)」、及び「**強化学習**(Reinforcement Learning)」に分類できる[8]。

教師あり学習は、その名の通り、教師データを用いて、学習主体（Learner）が解析・抽出した結果を評価し、学習主体の処理方式を改良する過程である。代表的な手法は「決定木(Decision Tree)、多層パーセプトロン(MLP: Multi-Layer Perceptron)、サポートベクトルマシン(SVM: Support Vector Machine)、単純ベイズ分類器(Naive Bayes Classifier)やベイジアンネットワーク(Bayesian Network)などが挙げられる。

一方、**教師なし学習**は、観測したサンプルデータに潜む規則や特徴を統計的な手法によって見出し、データのパターン認識やクラスタリングを行う。代表例として、k-means 法、主成分分析(PCA: Principle Component Analysis)、自己組織化マップ(SOM: Self-Organizing Map)などが良く知られる。

また、認知対象である「**環境**」に関する事前情報を必要とせず、学習主体が環境との相互作用と、その過程に伴う報酬や罰を通して、最適な「行動方策」を学習によって獲得する**強化学習**は、1980 年代から提案され、最適制御(Optimal Control)や自律ロボット(Autonomous Robot)の開発などの分野への応用が成功している[14]。更に近年、社会性を持つ群知能(Swarm Intelligence)の実現を目指すマルチエージェントシステム(MAS: Multi-Agent System)やスワームロボティックス(Swarm Robotics)の開発に、強化学習の導入が期待されている[4] -[13]。

本論文で提案する知的個体の「未知環境に適応する」能力は、自己組織化ファジィニューラルネットワーク(SOFNN: Self-Organizing Fuzzy Neural Network)を用いた強化学習

システムによって実現される。また、複数の知的個体の分散・群探索によって、それぞれの知的個体の単独学習・探索と比べ、より迅速な学習収束が可能であることを明らかにする。

強化学習の研究と問題点

一般的に、自律エージェントを構成する強化学習システムの基本的な要素は、観測できる環境情報である「状態(State)」、取り得る振る舞いである「行動（Action）」、目標を実現するための行動選択手法である「方策（Policy）」と、方策を改善することに用いられる環境による「報酬（Reward）」から構成される[14] [15]。**強化学習アルゴリズム**は、「学習主体であるエージェントの能動的な**探索行動**と、その行動結果によって環境から正または負の報酬を取得し、**方策を改善する**」という試行錯誤の繰り返しである。

エージェントの有限な取り得る行動によって、環境の状態が遷移する。未知環境の状態数が有限である場合、エージェントの行動決定過程はマルコフ性を持ち、行動を決定する方策が確率関数である場合、環境の状態遷移は、「**マルコフ決定過程**（MDP: Markov Decision Process）」となる[16] [17]。即ち、次の状態を決定するのは、現状態と現行動のみであり、以前の状態と行動の履歴に依存しない。また、観測方法によって、状態の観測が完全でない場合もしばしばある。その場合は、「**部分観測マルコフ決定過程**（POMDP: Partially Observable Markov Decision Process）」と呼ばれ、非完全知覚で観測した状態が同じでも、有限な決定過程の中で最適解に接近するため、異なる次状態へ遷移する行動方策が要求される。なお、状態と行動のみでなく、長期遅延報酬などの要素も状態遷移に影響する場合は、「セミマルコフ決定過程（SMDP: Semi-Markov Decision Process）」と呼ばれる。

強化学習は、ゴール指向型の能動的機械学習として主に以下の３つのアプローチがある[15][18]-[20]。

(i) 分類子システム(Classifier)と Profit-sharing 学習：
入力情報に応じて、報酬の高い行動を出力するルールをいかに発見・生成し、報酬に基づいてルールに信頼度（Credit）を与え、最適な時系列順のルール集合を決定する推論・帰納型学習方式である[21] [22]。

(ii) Actor-Critic を用いた TD 学習：
環境の状態の特徴に応じる行動価値を生成する Actor を構成し、それを用いて行動を選択する。また、状態の特徴を用いて状態価値を生成する Critic を構成し、その状態価値の変化である TD 誤差（Temporal Difference Error）に応じて行動の結果を評価し、行動価値関数を改善する試行錯誤型学習方式である[23]-[27]。

(iii) 動的計画法(DP: Dynamical Programming)を用いた TD 学習：
マルコフ決定過程(MDP)の環境で、Bellman 方程式の最適解を求め、状態価値関数や、状態―行動価値関数の変化量を行動選択方策の修正に反映する学習法であ

る。Q 学習[28]や Sarsa 学習[29]と呼ばれる学習アルゴリズムは最も洗練されたもので、最適制御などの分野でよく利用される。

しかし、自律ロボットが扱う実環境や、マルチエージェントシステム(MAS: Multi-agent System)などの大規模システムに対応するため、強化学習には以下の重要な課題が従来から指摘され、今日でも依然として挑戦的な課題となっている[18][19]。

(i) 観測データから特徴のある状態空間への推定：

強化学習は環境が未知と想定している。入力データをいかにクラスタリングして、入力の特徴を表現する状態空間を構成するかは、以前から研究され、多数のアプローチがある[29-52]。観測情報を分類することは、パターン認識やシステム同定・推定理論と本質的な違いはない。高次元入出力空間を持つシステムの「次元の呪い」がいかに解けるかは一つの重要な課題である。これまで、線形関数近似モデル[33]-[35]、ランダムタイリング[34]、適応共鳴理論(ART: Adaptive Resonance Theory)[36]、ファジィ推論[37]-[46]、ニューラルネットワーク(ANNs: Artificial Neural Networks) [14] [19] [47][48]、遺伝アルゴリズム(GA: Genetic Algorithm) [49]-[51]や免疫ネットワーク(AIN: Artificial Immune Network)[52]-[54]などの手法を用いた状態空間推定法が提案されている。

(ii) 状態空間の不完全性への対応：

強化学習の４つの要素である状態、方策、行動、報酬のうち、状態の認知に関する完全性が最も重要であるが、実環境の場合は、しばしば「不完全知覚」と呼ばれる問題が生じる。すなわち、「同じに見える状態」が実は異なる状態であり、その場合の方策は異なる行動を選択しなければならない。例えば、近傍観測の局所環境状態が同じでも、大局的に観測すれば異なる状態となる。また、短期的に観測した挙動が同じでも、長期的に見れば異なる挙動パターンになることもある。この場合の状態遷移過程は「部分観測マルコフ決定過程（POMDP: Partially Observable Markov Decision Process）」と呼ばれ、MDP を条件とする強化学習法をそのまま POMDP の問題に適用できない[4]-[13][55]-[59]。POMDP の環境での強化学習はこれまで「メモリによる状態の区別」を行う手法や、「適格度履歴（eligibility trace）」、「メモリレスの Profit Sharing」など意欲的なアプローチが提案されているが、問題の困難さより、今後より高性能・高効率の手法が期待される[59]。

(iii) マルチエージェントの学習：

環境が未知の上、動的であるマルチエージェント環境での強化学習は、近年、分散制御や群知能などの分野で活発に展開され、今後の発展が興味深く注目されている[4-13] [55]-[59]。エージェントがそれぞれ独立に行動・学習するため、状態の遷移が不確定であり、非 MDP となる。ただ、エージェントに共通な目標を定め、強化学習アルゴリズムを用いて、群行動を獲得させることによって、局所解への収束は期待できる。

本論文で提案する強化学習システム

本論文では、特に分散制御型システムの代表であるマルチエージェントシステム(MAS: Multi-Agent System)に注目し、複数の自律エージェントによる集団行動―群知能（Swarm Intelligence）を獲得するため、独自の自律エージェントを提案する。また、行動の目的があらかじめ指定される一般的なエージェントと比べ、以下の機能をより明確に実現できることを念頭に提案する自律エージェントを「知的個体（Intelligent Individual）」と呼ぶ：

(i)　能動的に環境を知覚できること；
(ii)　入力情報の特徴を抽出できること；
(iii)　自ら到達目標を定められること；
(iv)　最小コストで定めた目標に到達できること。

本論文では、まず、知的個体の内部モデルとして、自己組織化ファジィニューラルネットワークを用いた強化学習システムを提案する。また、学習則が異なる場合の強化学習システムをそれぞれ構築する。さらに、提案した内部モデルを持つ複数の知的個体が未知の探索環境に存在する場合、群れの形成および適応群行動の獲得について考案する。

ここで、知的個体とは、「未知環境に適応できる行動主体」であり、一般的な知的エージェントに比べ、より環境の観測・認知及び行動の学習能力が高い「インテリジェントシステム」のことを指す。また、内部モデルとは、状態の認知及びその認知機構の形成、方策構成および改善、探索(Exploration)と利用(Exploitation)を実現できる行動の出力、など複数の知的機能を実現する数理モデルである。

1.2　本研究の成果

本研究の成果は以下に挙げられる。

(i)　強化学習の状態推定問題に対し、自己組織化ファジィニューラルネットワーク(SOFNN: Self-Organizing Fuzzy Neural Network)を用いて対応し、観測データ駆動型の状態推定手法を提案した（第 2 章）。SOFNN は従来の線形関数近似器、ニューラルネットワーク、ランダムタイリングなどの状態推定手法の働きと同じであるが、離散的状態空間と連続的状態空間を共に自動的に構成することができる特徴を持つため、より優れた状態推定法である。

(ii)　従来の代表的な強化学習アルゴリズム Actor-Critic、Q-learning、Sarsa 学習をそれぞれ、(i)の SOFNN と融合し、ファジィニューラルネットワーク(FNN)を用いた 3 種類のニューロファジィ強化学習システム FAC、FQ と FS を提案した。また、未知環境での目標探索問題に対し、計算機シミュレーションを用いて、提案した各強化学習システムの有効性を確認することができた（第 3 章）。

(iii)　提案した各種の強化学習システムを知的個体の内部モデルとして、個体間の距離を適切に保つことを群学習に導入し、複数個体の適応行動を獲得することができ

た。群学習によって、多点探索することができ、単独学習に比べ、解への収束が大幅に加速した（第 3 章と第 4 章)。

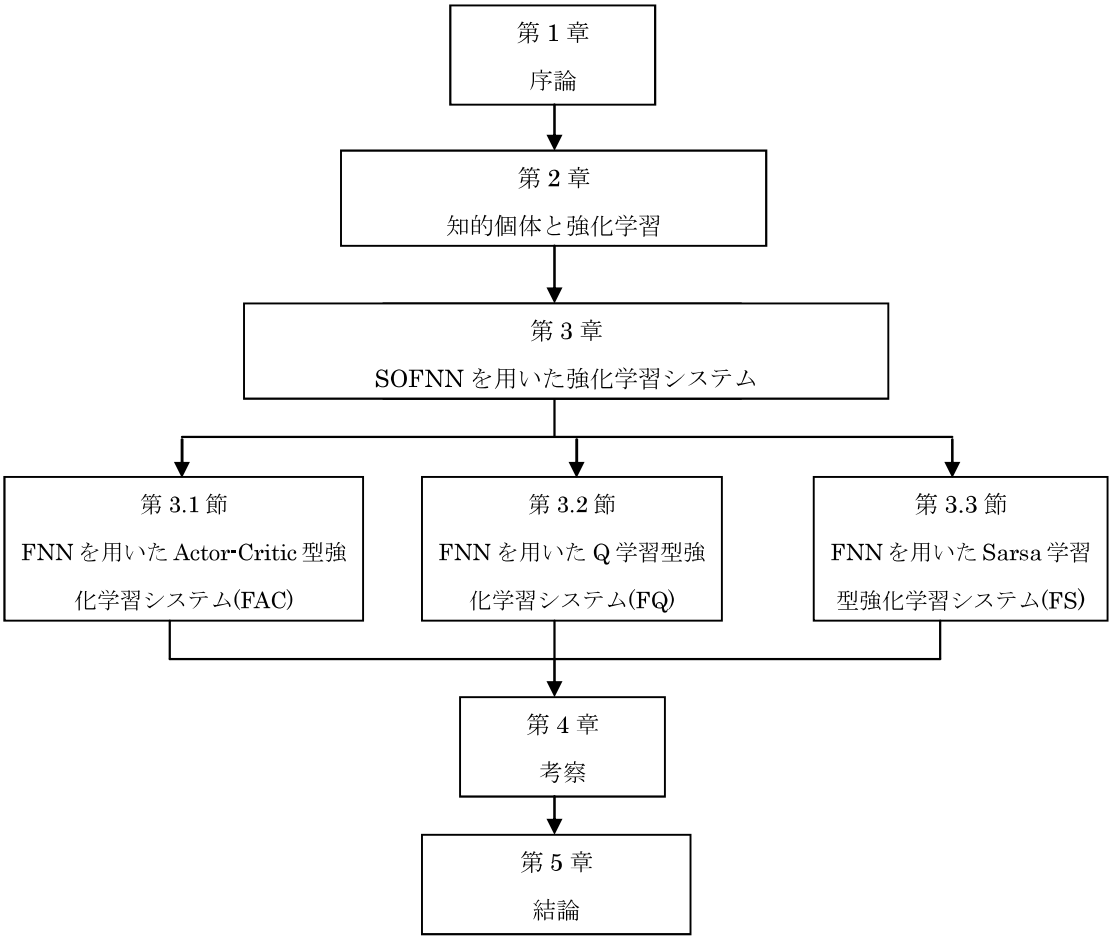


図 1.1 本論文の構成

1.3　本論文の構成

本論文の構成は、図 1.1 に示し、具体的には、以下の通りである。

第 1 章では、研究の背景と位置付け、及び論文の構成について述べる。

第 2 章では、知的個体の概念と機械学習の概要を述べ、本研究で用いられるファジィニューラルネットワーク及び強化学習について述べる。

第 3 章では、観測情報を分類するための自己組織化ファジィニューラルネットワーク（SOFNN）を用いた 3 種類の強化学習システムを提案する。ここで、「自己組織化」とは、入力データ駆動によるネットワークの自動形成を意味する。多次元入力空間による入力に対し、提案手法では、閾値制御及びルール生成規則によって、ファジィメンバーシップ関数や、ファジィルールを生成・結合し、ファジィネット（FN）を構成する。本章では、

まず、3.1 節では FN を用いた Actor-Critic 型強化学習システム（FAC）を提案する。

具体的には、SOFNN の出力に荷重を加え、ネットワークの一層とする状態価値を表す Critic モジュールと行動価値を表す Actor モジュールを可塑的に結合する。Actor の出力は、確率探索をもたらす行動選択方策関数に用いられ、知的個体が行動を選択し、出力する。行動の結果による環境の状態変化と、知的個体に返す報酬や罰を含む時間差分誤差（TD-error）を用いて、SOFNN と Critic・Actor の結合荷重を修正することよって、知的個体が価値の高い状態へ遷移するように、行動方策が修正される。座標情報（離散値及び連続値）を用いた有限マルコフ決定過程(MDP)を持つ目標探索問題のシミュレーションを行い、提案した FAC によって構成された知的個体が環境との相互作用の結果より、適切な行動を獲得できることが確認された。

次に、3.2 節では、FN を用いた Q 学習型強化学習システム（FQ）を提案する。システムの構成は先述の FAC と異なり、SOFNN の出力は、状態―行動価値を表す関数（Q 関数）と可塑的に結合する。行動選択方策関数は Q 値を用いて構成される。1989 年に Watkins により提案された一般的に利用されている Q 学習アルゴリズムと異なり、従来の「TD 誤差を直接 Q 値の修正に用いる」というアプローチの代わりに、TD 誤差を FAC の結合荷重の修正に導入し、システム（FQ）の出力改善につなげる。近傍情報しか得られない部分観測マルコフ決定過程(POMDP)を持つ目標探索問題のシミュレーションを通して、提案した FQ によって構成された知的個体が環境との相互作用の結果より、適切な行動を獲得できることが確認された。

最後に、3.3 節では、FN を用いた Sarsa 型強化学習システム（FS）を提案する。FS の構成は前節で提案した FQ と同じであるが、Q 学習の TD 誤差の計算法と異なり、1994 年に Rummery & Nirajan により提案された Sarsa 学習の TD 誤差を FS の結合荷重の修正に導入する。近傍情報しか得られない部分観測マルコフ決定過程(POMDP)を持つ目標探索問題のシミュレーションを通して、提案した FS によって構成された知的個体が環境との相互作用の結果より、適切な行動を獲得できることが確認された。

また、魚や鳥など生物の群行動の概念を未知環境における目標探索問題の解法に導入し、知的個体間の適切な距離を保つように、「離れすぎず近すぎず」行動の報酬を個体の価値関数に反映し、「群学習」によって、最適解、または準最適解をより早く発見することを図る。「群学習」の概念をそれぞれの提案システムに導入した場合と、他の個体との距離を考慮しない「単独学習」の場合の比較が、計算機シミュレーションの結果によって行われる。

第 4 章では、不完全観測環境（POMDP）における目標探索問題のシミュレーションを用いて、ランダム探索、従来の Q 学習法、従来の Sarsa 学習法、従来の SGA 学習法及び提案した FAC、FQ と FS のそれぞれの学習結果・学習性能の比較を行い、提案法の有効性を確認する。また、提案法におけるパラメータ設定方法について考察する。

第 5 章では、本論文のまとめと今後の課題について述べる。

1.4　本論文に関する補足説明

本論文の第 2 章に述べる自己組織化ファジィニューラルネットワーク(SOFNN)は、以下の査読付き国際会議・シンポジウム及び論文誌で、著者が発表した論文に提案・記述されている。

(i) The 35th ISCIE International Symposium on Stochastic Systems Theory and Its Applications (SSS ’03) [38];
(ii) The 2nd International Conference on Computational Intelligence, Robotics and Autonomous Systems (CIRAS ’03) [39];
(iii) The 27th Annual International Symposium on Forecasting (ISF 2007) [41];
(iv) 2008 IEEE World Congress on Computational Intelligence / International Joint Conference on Neural Networks (WCCI 2008 / IJCNN 2008) [85];
(v) 2008 International Conference on Intelligent Computing (ICIC 2008) [86];
(vi) International Journal of Intelligent Computing and Cybernetic (IJICC), Vol.2, No.4, pp.724-744, 2009 [87];
(vii) Journal of Circuits, Systems, and Computers (JCSC), Vol.18, No.8, pp.1517-1531, 2009 [88];
(viii) 2011 International Conference on Future Wireless Networks and Information Systems (ICFWI 2011) [92];
(ix) 電気学会論文誌Ｃ(IEEJ Transaction on EIS)，Vol.133, No.5, pp.1076-1085, 2013 [93].

また、本論文の第 3.1 節で提案する FNN を用いた Actor-Critic 型強化学習システム(FAC)は、上記(iv)-(vii)の発表論文[85]-[88]を基に、第 3.2 節で提案する FNN を用いた Q 学習型強化学習システム(FQ)は、上記(viii)の発表論文[92]を基に、第 3.3 節で提案する FNN を用いた Sarsa 学習型強化学習システム(FS)は、上記 (viii)及び(ix)の発表論文[92][93]を基にしている。

第 2 章

知的個体と強化学習

本研究の目的は、知能を持つ人工物を構築することよって、科学技術の発展及び人類の幸福に貢献することである。その「知能を持つ人工物」は、本論文で、「知的個体(Intelligent Individual)」と呼ぶ。「個体」と称する理由は、真の知能を持つ人工物であれば、他の人工物や実世界の環境と相互作用が必ず存在し、その「群れ」や「社会性」を念頭に置いているからである。なお、ここでの「人工物(Artifacts)」とは、一種または多種の機能を持つ数理モデル、ソフトウェア及びハードウェアを概に指す。

2.1　知的個体の構成

「知能(Intelligence)」の一般的な定義は、「環境に適応し、新しい問題の状況に対処する知的機能、能力」であり（『広辞苑』第 5 版）、または、「生物の適応形式の最高次の機能」である（『心理学辞典』、有斐閣）。さらに、工学的に「環境に適応して、自らを変化させ、あるいは適切な行動を選択する計算能力」という定義がある[9]。

本論文で提案する知的個体は、「能動的学習によって、未知環境に適応しながら、自ら行動の目的を発見し、その目的を実現するため、適切な行動を出力する」機能を有するインテリジェントシステムである。

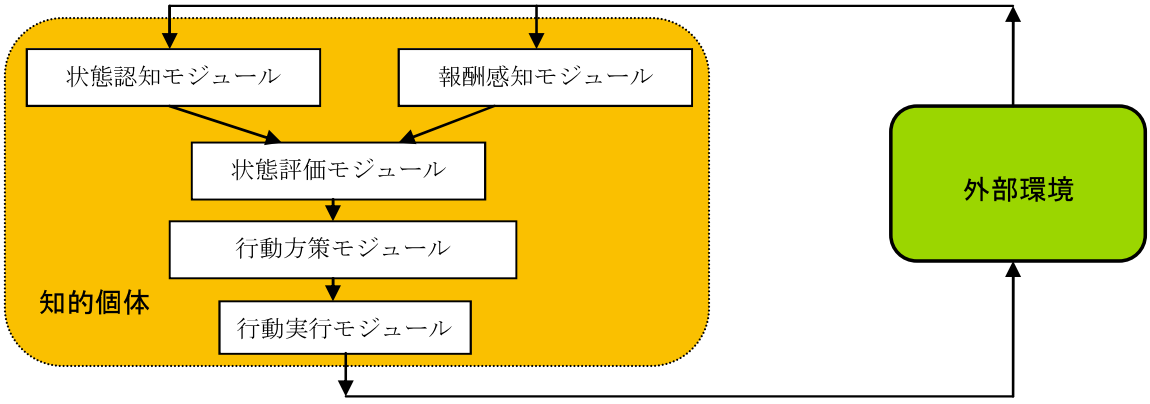


図 2.1 知的個体の構成及びその環境との相互作用

図 2.1 は知的個体の構成及びその環境との相互作用を示す。知的個体の環境認知・適応機能は、「状態認知」、「報酬感知」、「状態評価」、「行動方策」、「行動実行」という５つのモジュールによって実現され、それぞれのモジュールは、処理の流れによって結合され、「内部モデル」を構成している。

(i) 状態認知モジュールは、観測データに対し、自動的に離散、または、連続状態空間へのクラスタリングを実現する機能を持ち、自己組織化ファジィニューラルネットワーク（SOFNN: Soft-Organizing Fuzzy Neural Network）[38]-[46]によって構成される（第 2.2 節参照）;

(ii) 報酬感知モジュールは、異なる環境の状態が知的個体に対する「価値」となる報酬や罰（スカラー値）を受け取り、状態や行動の評価に用いられる（第 2.3 節、第 3 章参照）。

(iii) 状態評価モジュールは、状態遷移に伴う報酬及び報酬の和を計算し、最適な状態遷移過程を実現するため、各状態に価値を与える処理を行う（第 2.3 節、第 3 章参照）。

(iv) 行動方策モジュールは、学習主体の状態遷移を実現させるルールである。短期的に目前の最適状態に遷移させる「貪欲的(greedy)」な方策もあれば、確率分布関数などを用いる探索的な確率方策は、より長期的な報酬を獲得することを可能にする（第 2.3 節、第 3 章参照）。

(v) 行動実行モジュールは、行動方策によって、知的個体の取り得る行動を選択し、それを出力する（第 2.3 節、第 3 章参照）。また、行動方策に推奨された複数の行動ベクトルによる線形結合を行い、任意の連続行動を出力する（第 3 章参照）。

2.2　ファジィニューラルネットワーク

エアコンや洗濯機などの家電にも応用されているファジィ制御技術は、1990 年以降世界中に知れ渡っている。「ファジィ」とは「曖昧な」という意味であり、1965 年の L. A. Zadeh の論文“Fuzzy Sets” [96]でファジィ集合の概念が提案され、ファジィ推論、ファジィ制御、ファジィニューラルネットワークなどの技術・理論が迅速に発展し、広く応用されてきた[97]-[107]。

本節では、まず、ファジィ理論の基本概念を述べ、次に、教師学習機能を有する人工神経回路網であるニューラルネットワークについて簡単に紹介し、そして、ファジィ推論とニューラルネットワークの融合である自己組織化ファジィニューラルネットワーク（SOFNN: Self-Organizing Fuzzy Neural Network）を記述する。

2.2.1 ファジィ集合

全体集合を Ω とし、部分集合を $A \subset \Omega$ とする。Ω に属する要素 x が部分集合 A に属する度合い $\chi_A(x)$ はバイナリの値{0, 1}となる場合、その度合いは(2.1)式の「特性関数」で表され、Ω はクリスプ集合と呼ばれる。

$$\chi_A(x) = \begin{cases} 1 & if\ x \in A \\ 0 & otherwise \end{cases} \tag{2.1}$$

一方、要素 x が集合 A に属する度合いが[0, 1]区間の実数値で表現される場合、その度合い $\mu_A(x) \subset [0,1]$ は「メンバーシップ関数」と呼ばれ、集合 Ω と A は「ファジィ集合(Fuzzy Set)」と呼ばれる。

例えば、「c 付近の実数」というファジィ集合 B のメンバーシップ関数は(2.2)式のようなベル型（ガウシアン型）関数で表すことができる（図 2.2 参照）。

$$\mu_B(x) = \exp(-\frac{(x-c)^2}{\sigma^2}) \tag{2.2}$$

ここで、パラメータ $c \in R$ はガウシアン関数の中心、$\sigma > 0$ は広がりである。

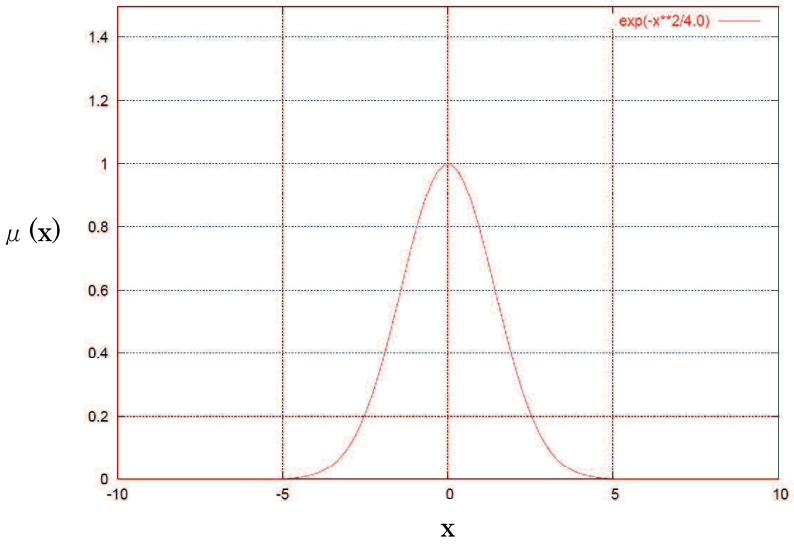


図 2.2 ガウシアン型メンバーシップ関数の形状（中心 c=0, 広がり σ=2）

2.2.2 ファジィ集合の合成

多次元変数の場合、それぞれの次元におけるファジィ集合に属する度合いは各次元のメンバーシップ関数の論理積、または、代数積に等しい。即ち、n 次元の変数 $\mathrm{x} \in X^n$ のメンバーシップ関数は以下のようになる。

$$\mu_{A_1 \times A_2 \times \cdots \times A_n}(\mathrm{x}) = \min\{\mu_{\mathrm{A}_1}(x_1),\quad \mu_{\mathrm{A}_2}(x_2), \cdots,\quad \mu_{\mathrm{A}_n}(x_n)\} \tag{2.3}$$

または、

$$\mu_{A_1 \times A_2 \times \cdots \times A_n}(\mathrm{x}) = \mu_{\mathrm{A}_1}(x_1) \cdot \mu_{\mathrm{A}_2}(x_2) \cdots \mu_{\mathrm{A}_n}(x_n) \tag{2.4}$$

但し、$x_1 \in X_1,\quad x_2 \in X_2, \cdots,\quad x_n \in X_n$ である。

2.2.3 ファジィ推論

「A なら B である」すなわち「IF A THEN B」という推論ルールが多数存在し、それらのルールによって、最終的な判断や行動を出力する論理システムは「プロダクションシステム」と呼ばれ、従来の人工知能分野では良く知られる。推論ルールの前件部(antecedent part)A、または後件部(consequent part)B にファジィ集合が含まれる場合は、その推論は「ファジィ推論(fuzzy inference)」と呼ばれ、推論ルールは「ファジィルール(fuzzy rule)」となる。

適応制御やシステム同定によく用いられる高木-菅野ファジィ推論モデル（TSK fuzzy model: Takagi-Sugeno-Kang fuzzy model）は以下のように構成されている[99]-[101]。

Rule R_i :

if x_1 *is* A_1 *and* x_2 *is* A_2 ... *and* x_n *is* A_n ,

then u *is* $y_i = f_i(x_1, x_2, \ldots x_n) = b_{i0} + b_{i1}x_1 + \ldots + b_{in}x_n$ (2.5)

ここで、$b_{i0}, b_{i1}, ..., b_{in}$ は実数パラメータである。

L 個のルールによる出力 y はメンバーシップ関数 $\mu_{A_{i1}}(x_1), \mu_{A_{i2}}(x_2), ..., \mu_{A_{in}}(x_n)$ を用いた重みつき y_i の線形結合関数となる。

$$y = \frac{\sum_{i=1}^{L} \tau_i y_i}{\sum_{i=1}^{L} \tau_i} \tag{2.6}$$

但し、$\tau_i = \mu_{A_{i1}}(x_1) \times \mu_{A_{i2}}(x_2) \times \ldots \times \mu_{A_{in}}(x_n)$ である。

2.2.4 ニューラルネットワーク

神経細胞の膜電位の変化を入出力関係のある数式でモデル化し、それらを複数個結合して構成した人工神経回路網であるニューラルネットワーク(ANN: Artificial Neural Network)は、1940 年代から研究され、1986 年の多層パーセプトロン(MLP: Multi-Layer Perceptron)の出現と共に計算知能の代表的な数理モデルとなり、パターン認識、システム同定、関数近似、予測、知的制御など多くの分野で応用されている[97] [98]。

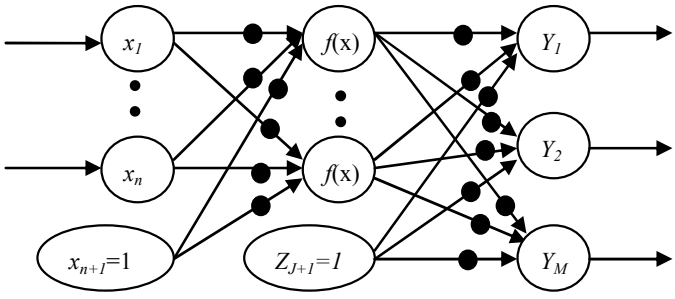


図 2.3 多層パーセプトロン(MLP)

一般的に、MLP は、同じ層のユニット（ニューロン）の結合がなく、層間だけが重みつきの単方向の結合があり、各ユニットは Sigmoid 関数や Gaussian 関数などの非線形関数を

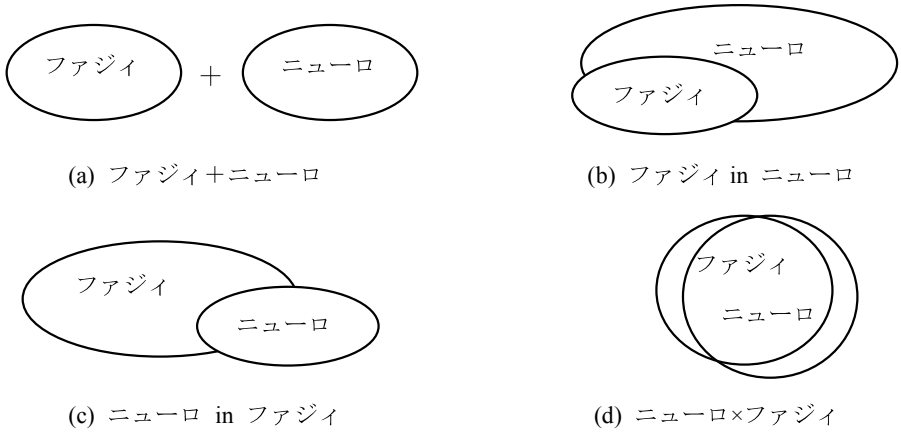
用いる層状ニューラルネットワークを指す（図 2.3 参照）。図 2.3 において、n 次元入力 $\mathbf{x}(x_1, x_2, \ldots, x_n)$ と常に 1 の値を取るバイアスユニット x_{n+1} が重み v_{ji} で中間ユニット $z_j = f(\mathrm{x}) = f(\sum_{i=1}^{n+1} v_{ji} x_i), j = 1, 2, ..., J$ と結合し、また、中間ユニットは z_j 出力層のユニット y_m と重み $w_{mj}, m = 1, 2, ..., M$ で結合し、MLP の出力は $y_m = f(z_j) = f(\sum_{j=1}^{J+1} w_{mj} z_j)$ となる。

ユニット間の結合荷重 v_{ji}, w_{mj}（図 2.3 中の●）は任意の初期値を持ち、教師信号との誤差を用いる更新則によって、MLP は任意の関数を近似することができ、また、多次元の入力データのパターンを分類・認識することができる。その結合荷重の修正プロセスは「学習 (learning or training)」と呼ばれる。

2.2.5 ファジィとニューラルネットワークの融合

2.2.1 節に述べられたファジィ推論は入力パターンの曖昧さを評価し、その評価結果に従って、適切な出力を定める知的情報処理のプロセスである一方、ファジィ集合を評価するメンバーシップ関数の設計や、ファジィ推論におけるファジィルールの決定法は人工的な事前設計が必要となる。また、前節で述べたニューラルネットワークは入出力の関係をユニット間の結合荷重の修正によって学習し、知識を記憶することができる。ファジィ技術とニューラルネットワーク技術の長短を考慮し、より適応能力の高い知的情報処理システムを構成するため、両者を融合した「ファジィニューラルネットワーク（Fuzzy neural networks）」、或いは「ファジィニューラルネットワーク」や「ファジィニューロ」手法が近年多く提案されている[97][98]。

図 2.4 にファジィとニューラルネットワークの 4 種類の大まかな融合形式を示す（「形態による分類」[97]）。



(a) ファジィ＋ニューロ　(b) ファジィ in ニューロ

(c) ニューロ in ファジィ　(d) ニューロ×ファジィ

図 2.4　ファジィとニューラルネットワークの融合[97]

本論文では、2.1 節で述べた「知的個体」（図 2.1 参照）を構成する際に、次節で述べる自己組織化ファジィニューラルネットワーク(SOFNN)を用いる。知的個体の「状態認知モジュール」は、ファジィ推論システムによって構築され、「状態評価モジュール」及び「方策決定モジュール」は、重み付きの結合を持つニューラルネットワークによって構成されるため、相乗効果のある図 2.4(d)の「ニューロ×ファジィ」型ニューロファジィシステムに属するであろう。

2.3　自己組織化ファジィニューラルネットワーク(SOFNN)

「環境の状態を観測する」ことは、外部の刺激信号を異なるパターンに分類し、認知することである。自律移動ロボットの場合は、視覚・聴覚・触覚・味覚・嗅覚などのセンサーを通して、外部状態、または自身の内部状態を把握する。

本論文で提案する知的個体は、その自律移動ロボットの知的情報処理システムとして利用されることを意識し、「自己組織化ファジィニューラルネットワーク（SOFNN: Self-Organizing Fuzzy Neural Network）を用いて、外部環境の状態分類・状態認知機能を実現する。

SOFNN は入力データに対して、ファジィ推論を行い、その推論結果によって、ファジィ集合を定義するメンバーシップ関数の増減及びファジィルールの増減を自動的に実現するデータ駆動型ファジィシステムであり、これまで、時系列予測システム[38][39][41][42]、制御システム[43]-[46]、強化学習システム[85]-[88][92]-[94]などの知的システムに導入されている。

ある時刻（ステップ）t において、環境状態に対し、入力ベクトル $\mathbf{x}(t)=(x_1(t),\ldots,x_i(t),\ldots,x_n(t))$ を観測した場合、それに対するファジィ推論の k 番目ファジィルール $R^k\,(k=1,2,\ldots,K_t)$ を(2.7)式のように定義する。

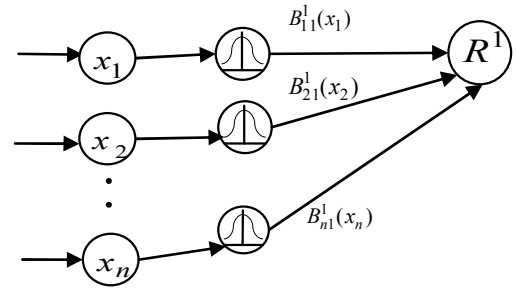
$$R^k : if\left(x_1(t)\ is\ B^k_{1m_1},\ldots,x_i(t)\ is\ B^k_{im_i},\ \ldots,\ x_n(t)\ is\ B^k_{nm_n}\right)\ then$$

$$\phi^k_t(\mathrm{x}(t))=\prod_{i=1}^{n}B^k_{im_i}\left(x_i(t)\right). \tag{2.7}$$

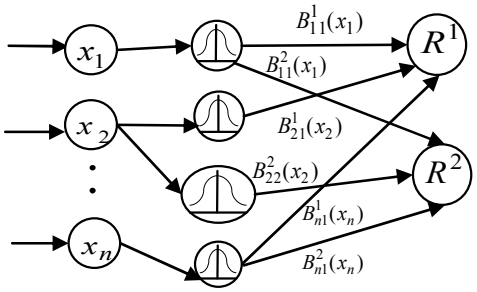
ここで、$B^k_{im_i}(x_i(t))$ はファジィ集合のメンバーシップ関数；i は入力ユニットの番号、m_i は i 番目の入力 $x_i(t)$ に関するメンバーシップ関数の番号である（$i=1,2,\ldots,n;\ m_i=1,2,\ldots,M_i(t)$）。$\phi^k_t(\mathbf{x}(t))$ はファジィルール R^k 後件部の出力（適合度）である（$k=1,2,\ldots,K_t$）。なお、$B^k_{im_i}(x_i(t))$ は(2.8)式のガウス動径基底関数を用いる(図 3.1 参照)。

$$B_{im_i}^k\left(x_i(t)\right)=\exp\left\{-\frac{1}{2}\left(\frac{x_i(t)-c_{im_i}^k}{v_{im_i}^k}\right)^2\right\}. \tag{2.8}$$

ここで，$c_{im_i}^k$ と $v_{im_i}^k$ はそれぞれ i 番目の入力 $x_i(t)$ に対する、k 番目のルールに接続するメンバーシップ関数 $B_{im_i}^k$ の中心と広がりである。



(a) 最初の入力を受けた場合（t=1）



(b) 新しいメンバーシップ関数及びファジィルールの増加

図 2.5　SOFNN におけるメンバーシップ関数とファジィルールの生成

SOFNN は入力データの駆動によってメンバーシップ関数やファジィルールが自己増減できるように設計される：

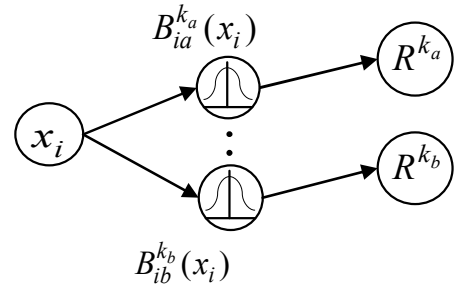
各入力ベクトルに関するメンバーシップ関数 $B_{im_i}^k(x_i(t))$ の数 $M_i(t)$ およびファジィルール $\phi_t^k(\mathbf{x}(t))$ の数 K_t は、0 個から始まり、未知の状態を観測する度に、それぞれ、自己増殖する。具体的には、$t=0$ のとき、ファジィ集合は存在せず、$t=1$ になると、各入力ベクトルに対し、一つ目のメンバーシップ関数が生成される。すなわち、入力 $x_i(t)$ に対応するファジィ集合の個数 $m_i=1$ で、メンバーシップ関数 $B_{i1}^1(x_i(t))$ の中心と広がりは $c_{i1}^1=x_i(t), v_{i1}^1=\sigma$ として、メンバーシップ関数 B_{i1}^1 とファジィルール R^1 が構成される（$i=1,2,...,n$）（図 2.5(a)参照）。ここで、σ は初期設定の正のパラメータである。

$t=2$ 以降、観測ベクトルの i 番目の要素に対するメンバーシップ関数を追加するための閾値を F とし、次式が成り立つ場合、新しいメンバーシップ関数とそれに対応する新しいルールを追加する（図 2.5(b)参照）。

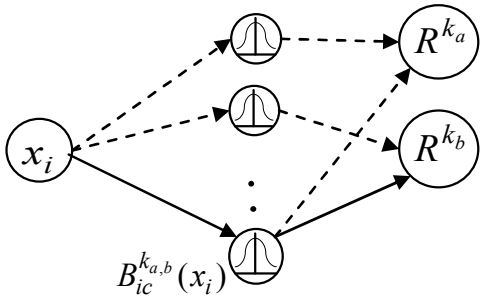
$$
\begin{aligned}
&if\ \max B_{im_i}^{k}\left(x_i(t)\right) < F \qquad \left(m_i = 1,2,\ldots,M_i(t)\right)\\
&\qquad\qquad then\\
&\qquad\qquad M_i(t) \leftarrow M_i(t)+1,\\
&\qquad\qquad K_t \leftarrow K_t + 1.
\end{aligned}
\tag{2.9}
$$

追加するメンバーシップ関数の中心 $c_{im_i}^{k}$ は観測ベクトルの i 番目の要素 $x_i(t)$ の値、広がり $v_{im_i}^{k}$ は $t=1$ と同様なパラメータ σ とする。

(2.9)式が成立すると、i 番目以外の他の入力のメンバーシップ関数は、それぞれの入力に対する最も大きい出力を持つメンバーシップ関数を選び、それらを新たに作成したルール $k=K_t$ へ接続する。結果として、一つの入力に関するメンバーシップ関数が複数生成される可能性があり、それらのメンバーシップ関数はそれぞれ異なるファジィルールに接続することになる。以上に述べた新規のメンバーシップ関数と新規のファジィルールの生成過程を図 2.6 に示す。



(a) ファジィ集合統合前



(b) ファジィ集合統合後

図 2.6　SOFNN におけるメンバーシップ関数の統合

また、入力データが新たに観測された時、i 番目入力要素 $x_i(t)$ のファジィ集合 $M_i(t)$ 個のうち、二つのメンバーシップ関数が似ている（中心と広がりが似かよった値となる）ことが考えられる。ここでは、以下の条件が満足されると、類似しているとし、その二つのメンバーシップ関数を一つの新しいメンバーシップ関数として統合する。

$$
\begin{aligned}
&\textit{if } \max B_{ia_i}^{k_a}(x_i(t)) > F^{\perp} \ \textit{and} \ \max B_{ib_i}^{k_b}(x_i(t)) > F^{\perp}, \\
&\qquad \forall a_i, b_i \in \{1,2,\ldots,M_i(t)\}, \forall k_a, k_b \in \{1,2,\ldots,K_t\} \\
&\qquad\qquad \textit{then} \\
&\qquad\qquad M_i(t) \leftarrow M_i(t) - 1, \\
&\qquad\qquad c_{ic_i}^{k_{a,b}} = \frac{1}{2}(c_{ia_i}^{k_a} + c_{ib_i}^{k_b}), \\
&\qquad\qquad v_{ic_i}^{k_{a,b}} = \frac{1}{2}(v_{ia_i}^{k_a} + v_{ib_i}^{k_b}).
\end{aligned}
\tag{2.10}
$$

ここで、メンバーシップ関数 $B_{ia}^{k_a}(x_i(t))$ と $B_{ib}^{k_b}(x_i(t))$ のグレードはともに十分高い場合（閾値 $F^{\perp}$ を超える）を考え、それらの類似を認め、$B_{ia}^{k_a}(x_i(t))$ と $B_{ib}^{k_b}(x_i(t))$ を統合し、新たなメンバーシップ関数 $B_{ic}^{k_{a,b}}(x_i(t))$ とする。また、$B_{ic}^{k_{a,b}}(x_i(t))$ のルール接続先は、$B_{ia}^{k_a}(x_i(t))$ と $B_{ib}^{k_b}(x_i(t))$ がそれぞれ接続していた 2 本のルール k_a, k_b となる（図 2.6 参照）。

なお、ファジィ集合の統合によって、ファジィルールが完全一致することも有り得る。すなわち、二つのファジィルールが全く同じ n 個の入力ファジィ集合を使用する場合が生じる。その場合は二つのファジィルールを一つに見なされ、非ファジ化処理する場合（次章参照）、それらのルールの重みの平均値を使用する。

本節で述べた自己組織化ファジィニューラルネットワーク SOFNN が環境の状態を認知するモジュールとして知的個体の内部モデルに導入する場合（第 3 章）、ファジィネット(FN)と略称する。

2.4　強化学習

人工知能分野に属する機械学習(machine learning)は、サンプルデータから、知識やルールを抽出し、予測や制御などの目的に役立つ計算的理論である。第 1.1 節に述べたように、機械学習は、計算的な処理の方式によって、概ね「教師あり学習」、「教師なし学習」及び「強化学習」に分類できるが[8]、学習システム自体は、いずれの学習方式でも、「学習データ」、「学習モデル」と「学習アルゴリズム」という三つの要素から構成される[60]。

また、学習モデルは、(i) 統計的確率モデル；(ii)ネットワークなどからなる構造を持つ学習モデルに分けられる。前者は、最尤推定法などの学習アルゴリズムを用いて、データの確率分布を表すモデルのパラメータを求める。後者は、人工神経回路網(ANNs: Artificial Neural Networks)や、ベイジアンネットワーク(BNs: Baysian Networks)、隠れマルコフモデル(HMMs: Hidden Markov Models)などを用いて、学習モデルを構築する。

本論文は、知的個体の内部モデルを開発するため、学習主体の能動性と自律性が発揮できる強化学習方式を用いる。

まず、未知の入力データの特徴を抽出するため、ファジィ推論とニューラルネットワークを融合したファジィニューラルネットワークを構築し、それを「学習モデル」として、新たな強化学習システムを提案する。提案モデルはユニットの結合によるネットワークで構成されているため、前述の「構造を持つ学習モデル」に属するが、強化学習アルゴリズムの「試行錯誤の反復」によってネットワークの結合荷重を修正するため、「統計的確率モ

デル」の特徴もある。

「学習データ」に関しては、知的個体が観測した環境の情報をニューロファジィネットによって、入力データの規則を抽出し、特徴によって入力データが分類され、学習システムへの入力となる状態空間が構成される(第 3 章参照)。

そして、「学習アルゴリズム」に関しては、これまでの強化学習の成果を基にして、知的個体の内部モデルのパラメータ学習のため、Actor-Critic 型 TD 学習（第 3.1 節参照)、Q-learning 型 TD 学習（第 3.2 節参照)、及び Sarsa 型 TD 学習アルゴリズム(第 3.3 節参照)を用いる。

2.4.1 強化学習の概念

機械学習分野に属する強化学習の研究は 1950 年代から始まり、これまで、数多くの手法が提案されているが(1.1 節参照)、マルチエージェント系の強化学習法について、状態の不完全性と不確定性から、最適解への接近は依然挑戦的な課題とされている。これに対し、1990 年代以降、概して、ルールの強度や状態遷移の履歴を重視する分類子システムと、状態の価値、または状態に応じる行動の価値の変化に注目する TD 学習法の二種類のアプローチに対して、それぞれ鋭意に研究が進められている[4]-[13]。本論文は自律ロボットの実現を念頭にし、「知的個体」のための学習システムを開発するため、オンライン性の高い TD 学習法を重点的に紹介する。

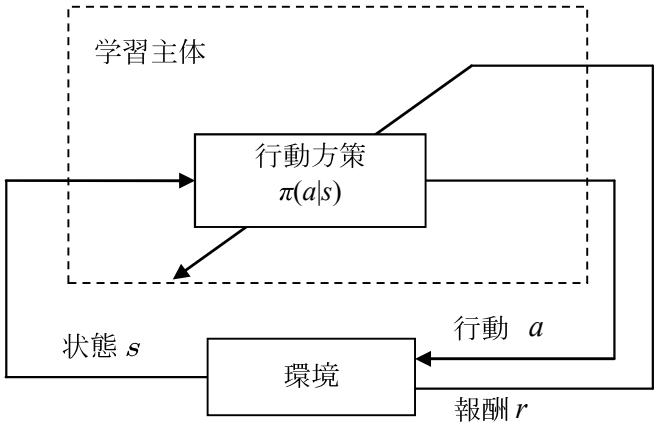


図 2.7 学習主体と環境の相互作用

強化学習(Reinforcement Learning)は、学習主体（Learner、エージェント、知的エージェント、自律エージェント、知的個体、強化学習システム、インテリジェンスシステム、自律ロボットなどの人工物を指す）と、学習主体に置かれた環境との相互作用によって、目標指向型の機械学習である。ここで、目標指向型とは、状態の遷移過程伴う報酬の獲得を目標状態に達成することによって最大にすることである。一般的に、状態の数が有限であることが想定され、終端状態に到達する場合は他の状態に到達する場合より大きい正の報酬が与えられ、その終端状態は、学習主体の最終タスク―目標となる。

図 2.7 は最も単純な学習主体と環境の相互作用を示す。学習主体と環境の相互作用とは、以下の過程の繰り返しを意味する。

【強化学習アルゴリズム】
Step 1. 学習主体は環境から状態 s を観測し、終端状態であれば終了する；
Step 2. 状態 s に対して、行動方策 $\pi(a|s)$によって、行動 a を出力する；
Step 3. 状態は s'に遷移し、環境から報酬また罰 r を得る。
Step 4. 報酬 r を用いて行動方策 $\pi(a|s)$を修正し、step 1 に戻る;

即ち、学習主体が、一連の状態の遷移を実現する「最適な行動方策」を見つければ、任意の状態から目標状態への遷移が実現でき、「最大報酬」を得ることができる。従って、強化学習の「学習」、或いは「学習アルゴリズム」は、Step 2 のより良い行動方策 π をいかに獲得するかのプロセスになる。

「強化（Reinforcement ）」は動物の行動分析に発端した用語で、ネズミなどの動物の自発的な行動の中、外部の「強化子（Reinforcer)」または「報酬（Reward)」の働きによって、良い結果の行動の生起度が高まる手続きを指す[10]。

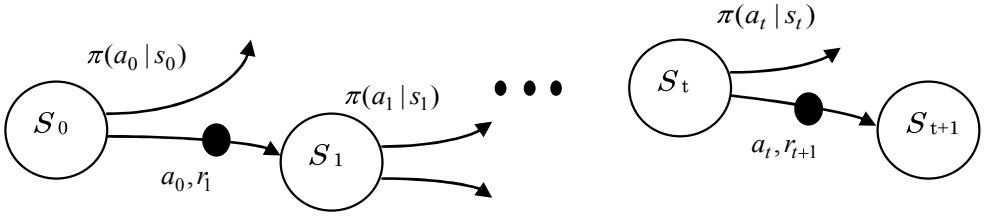


図 2.8 マルコフ決定過程下の状態遷移

2.4.2 マルコフ決定過程（MDP）と部分観測マルコフ決定過程（POMDP）

上述の強化学習問題を「マルコフ決定過程」で記述すると、以下のようになる。

離散時間のマルコフ決定過程(MDP: Markov Decision Process)は、状態集合 S と行動集合 A によって構成される。現在時刻 t の状態 $s_t \in S$ から次の状態 $s_{t+1} \in S$ への遷移は、行動方策（Policy）$\pi(a_t|s_t)$ に従って選択された $a_t \in A$ によって実現し、その遷移確率は $P_{ss'}^{a} \equiv Pr(s_{t+1} = s' | s_t = s, a_t = a)$ とする。状態遷移の「マルコフ過程(Markov Process)」は「次状態が現状態のみに依存する」と定義されているが、MDP は「次状態が現状態及び現行動によって出現する」遷移過程である。すなわち、

$$Pr(s_{t+1} | s_t, a_t) = Pr(s_{t+1} | s_t, a_t, s_{t-1}, a_{t-1}, \cdots, s_1, a_1, s_0, a_0) \tag{2.11}$$

が成り立つ。MDP 下の状態遷移を図 2.8 に示す。図 2.8 において、円は状態、矢印線は行動（遷移方向)、黒い点は行動が実行され、状態の遷移に伴い、得られる報酬を示している。

なお、不完全知覚によって、学習主体が観測した状態が部分的で、同じ状態とみなしても、実際は異なる状態に対し、同じ行動を出力してしまい、状態の遷移が実行される確率過程は「部分観測マルコフ決定過程(POMDP: Partially Observable Markov Decision Process)」と呼ばれる。

MDP と POMDP の比較を図 2.9 に示す。○で示される学習主体が（上, 下, 左, 右）の４方向（次元）の近傍環境を観測し(図 2.9 (a))、スタート位置 S から終了位置の G へ、トンネル型の簡単な迷路を探索する(図 2.9(a) (b))。白いマスが通路で、その値を 0 とし、黒いマスは観測値を 1 とする壁である。

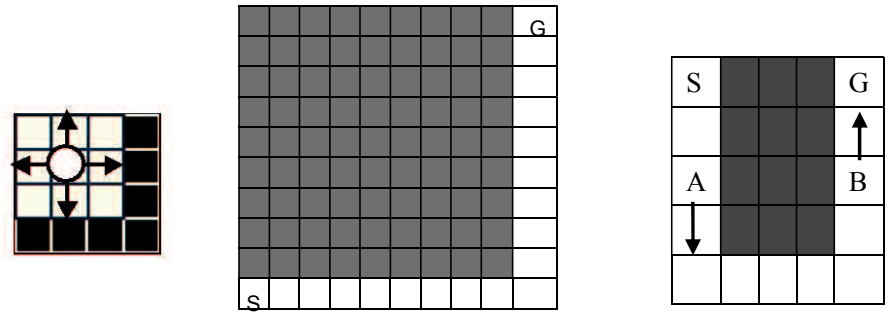
４次元の観測状態空間において、図 2.9(a)の学習主体の状態は(0, 0, 0, 0)となり、図 2.9(b)の状態数は以下の 5 個ある。

(1, 1, 1, 0)、(1, 1, 0, 0)、(0, 1, 0, 1)、(0, 0, 1, 1)、(1, 0, 1, 1)。

また、図 2.9 (c)の状態数は以下の 5 個になる。

(1, 0, 1, 1)、(0, 0, 1, 1)、(0, 1, 1, 0)、(1, 1, 0, 0)、(0, 1, 0, 1)。

図 2.9(b)のすべての状態に対し、「右へ」或いは「上へ」の最適行動が一つしかないため、MDP 環境となる。一方、図 2.9(c)の A 場所と B 場所の近傍観測状態は、いずれも(1, 1, 0, 0)であるが、目標の G 点に到達するため、それぞれの場所において、A 点の場合「下へ」、B 店の場合「上へ」となる異なる最適行動が要求されるため、図 2.9(c)の迷路は POMDP 環境となる。



(a) 観測状態　　(b)　MDP 環境　　(c) POMDP 環境

図 2.9 MDP と POMDP の比較

2.4.3 状態―行動価値関数

状態遷移に伴って、学習主体が報酬 $R(s_t,a_t;s_{t+1})$ を得るとし、MDP 下の任意の状態 $s\in S$ に得られる報酬の総和は、以下の状態―行動価値関数 $Q^{\pi}(s,a)$ で表す。

$$\begin{aligned} Q^{\pi}(s,a) &\equiv \mathop{\mathbf{E}}_{P^{a}_{ss'},\pi}\{\sum_{t=0}^{\infty}\gamma^{t}R(s_t,a_t;s_{t+1})\,|\,s_0=s,a_0=a\} \\ &= \mathop{\mathbf{E}}_{P^{a}_{ss'},\pi}\{R(s_0,a_0;s_1)+\sum_{t=1}^{\infty}\gamma^{t}R(s_t,a_t;s_{t+1})\,|\,s_0=s,a_0=a\} \end{aligned} \tag{2.12}$$

ここで、$\underset{P^a_{ss'},\pi}{\mathbf{E}}\{\cdot\}$ は、方策 π による定常確率 $P^a_{ss'}$ を持つ状態遷移過程における（報酬）期待値を表し，$\gamma(0 \le \gamma \le 1)$ は割引率である。

$P^a_{ss'} \equiv Pr(s_{t+1} = s' | s_t = s, a_t = a)$ のため、(2.12)式は

$$\begin{aligned} Q^\pi(s,a) &= \underset{Pr(s'|s,a)}{\mathbf{E}} \underset{\pi(a'|s')}{\mathbf{E}} \{R(s,a;s') + \gamma Q^\pi(s',a')\} \\ &= \underset{Pr(s'|s,a)}{\mathbf{E}} \{R(s,a;s')\} + \gamma \underset{Pr(s'|s,a)}{\mathbf{E}} \underset{\pi(a'|s')}{\mathbf{E}} \{Q^\pi(s',a')\} \end{aligned} \tag{2.13}$$

と書き直せる。

従って、強化学習問題の目的は、「状態―行動価値関数 $Q^\pi(s, a)$を最大にする最適な行動方策 $\pi^*(a|s)$を求める」ことになる。

$$\begin{aligned} &\pi^*(a \mid s): \\ &Q^*(s,a) = \max_\pi Q^\pi(s,a) \\ &\qquad = \max_\pi \underset{Pr(s'|s,a)}{\mathbf{E}} \{R(s,a;s' + \gamma Q^\pi(s',a')\}, \forall s \in S, \forall a \in A \end{aligned} \tag{2.14}$$

なお、過去（または将来）の状態履歴による収益をすべて考慮した目的関数(2.12)式は、最適解を求めるアルゴリズムの動的計画法(Dynamic Programming)では、Bellman 方程式と呼ばれる[15]-[17]。

2.4.4 行動方策

状態遷移の確率は、行動方策によって決定されるが、行動方策は、状態―行動価値関数を用いて学習主体の適切な行動を選択する。行動を決定する方策は、以下の 3 種類の確率分布関数が良く用いられるが、後者の二つは行動選択にランダム性があり、報酬を得ることが可能な知識利用(Exploitation)と環境の探索(Exploration)を共に実行することになる。

【行動方策 1】MaxQ 法：

任意の状態 s において、最大の状態―行動価値関数 $Q(s, a)$を持つ行動 a を選択する。

$$\pi(a \mid s) = \begin{cases} 1 & if\ a = \arg\max_{a \in A} Q(s,a), \\ 0 & otherwise. \end{cases} \tag{2.15}$$

【行動方策 2】ε-greedy 法：

最大の状態―行動価値関数 $Q(s, a)$を持たないすべての行動の選択確率は小さい乱数 0<ε<<1 で決め、最大の $Q(s, a)$値を持つ行動は比較的大きい確率で選択する。

$$\pi(a \mid s) = \begin{cases} 1 - \varepsilon + \dfrac{\varepsilon}{|\mathbf{A}|} & if\ a = \arg\max_{a \in A} Q(s,a), \\ \dfrac{\varepsilon}{|\mathbf{A}|} & otherwise. \end{cases} \tag{2.16}$$

【行動方策 3】Softmax 法：

変数 $Q(s, a)$のボルツマン分布に従って、行動 a を選択する。パラメータ $T>0$ は温度定数と呼ばれ、T の値が大きいほど、行動を選択するランダム性が高まる。

$$\pi(a|s)=\frac{\exp(Q(s,a)/T)}{\sum_{b\in A}\exp(Q(s,b)/T)} \tag{2.17}$$

また、行動方策を改善することは、以下の式で表すことができる。

$$Q^{\pi'}(s,a_{\pi'})>Q^{\pi}(s,a_{\pi}) \tag{2.18}$$

ここで、π' は π の改善で、$a_{\pi'}$ は π' によって選択された行動で、a_{π} は π によって選択された行動である。

以上の行動方策 $\pi(a|s)$ によって、学習主体は常に状態―行動価値関数 $Q^{\pi}(s, a)$の大きい、状態に適する行動を選択することになり、また、MDP が繰り返される場合、行動方策が改善される。

行動方策を改善する「学習」方法は、これまで、学習のタイミングによって、「動的計画法」、「モンテカルロ法」、「TD 学習法」などの提案があり、これらの学習法について第 2.4.5 節～第 2.4.7 節で述べる。

2.4.5 動的計画法

(2.12)式と(2.13)式の状態―行動価値関数 $Q^{\pi}(s, a)$の値は、報酬 $R(s, a; s')$と将来の報酬 $Q^{\pi}(s,' a')$の期待値によって与えられる。すなわち、$Q^{\pi}(s, a)$は状態 s に到達する回数は（無限大に近い）十分大きい場合、常に行動 a を選択し、収束した平均報酬値を表す。(2.14)式の最適状態―行動価値関数を求める Bellman 最適方程式は以下となる。

$$\begin{aligned}&\pi^*(a|s):\\ &Q^*(s,a)=\max_{\pi}\mathop{\mathrm{E}}_{Pr(s'|s,a)}\{R(s,a;s')+\gamma Q^{\pi}(s',a')\}\\ &\quad=\mathop{\mathrm{E}}_{Pr(s'|s,a)}\{R(s,a;s')+\gamma\max Q^*(s',a')\}\\ &\quad=\sum_{s'}P^a_{ss'}[R(s,a;s')+\gamma\max Q^*(s',a')],\forall s\in S,\forall a\in A\end{aligned} \tag{2.19}$$

すべての状態 s から s'への遷移確率 $P_{ss'}$は既知の場合、(2.9)式の解は部分問題の解によって与えられ、このような強化学習法は動的計画法(DP: Dynamic Programming)と呼ばれる。

しかし、強化学習の環境は一般的に未知であるため、DP 法の利用は限られる。

2.4.6 モンテカルロ法

MPD の状態数が有限である場合、初期状態から、最終状態に到達する遷移過程は 1 エピソード（Episode）と呼び、エピソードが反復すれば、(2.18)式の行動方策改善が実行される。

なぜならば、同じ状態 s に到達する際に、(2.15)式～(2.17)式のいずれが実行され、より大きい $Q^{\pi}(s,' a')$を得る行動 a が選択されるからである。エピソードの回数が十分大きければ、状態―行動価値関数の平均値への収束が実現できる。

前節に述べた DP 法における状態遷移確率を直接用いず、MDP の反復によって初期の行動方策を改善する手法は、強化学習のモンテカルロ法(Monte Carlo Method)と呼ばれる。モンテカルロ法はオフライン型の学習で、最適解への収束性が保証されるが、高次元の状態・行動空間での十分な繰り返しによって最適行動方策を求めるため、その計算量は膨大になる。

2.4.7 TD 学習法

前節に述べたモンテカルロ法のすべての状態遷移が終了し、エピソードごとに方策を評価・改善する学習アルゴリズムより、即時に状態―行動価値関数を修正し、行動方策を改善する TD 学習法（TD-learning: Temporal Difference Learning）の方がオンライン性が高く、また、その実行効率が高いことは明白である。

ある時刻 t における状態 s の期待価値 $V(s_t)$と、次の時刻 t+1 の状態の期待価値 $V(s_{t+1})$との間に、報酬 r_{t+1} 分の差があり、行動方策 $\pi(a|s)$の改善は、行動 a を選択することによって、新たな状態価値$(r_{t+1}+\gamma V(s_{t+1}))$を予測する意味を持つ。TD また TD 誤差(TD-)))Error ε は、その予測誤差を指す。

$$V^{\pi}(s_t) \equiv \underset{P^{a}_{ss'},\pi}{\mathbf{E}} \{R(s,a;s') \mid s = s_t\} \tag{2.20}$$

$$\varepsilon = r_{t+1} + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t) \tag{2.21}$$

TD 誤差を用いた価値関数の学習則は以下となる。

$$\begin{aligned} V^{\pi}(s_t) &\leftarrow (1-\alpha)V^{\pi}(s_t) + \alpha(r_{t+1} + \gamma V^{\pi}(s_{t+1})) \\ &= V^{\pi}(s_t) + \alpha\varepsilon \end{aligned} \tag{2.22}$$

ただし、$0<\alpha<1$は学習率である。

定常確率を持つ行動方策によって、行動の選択が行われるならば、状態の遷移は定常過程となるため、価値関数$V^{\pi}(s_t)$ と状態―行動価値関数$Q^{\pi}(s,a)$ の関係は以下となる。

$$Q^{\pi}(s,a) = \mathbf{E}\{V(s_t) \mid s_t = s, a_t = a, \pi\} \tag{2.23}$$

状態―行動価値関数を用いた TD 学習は以下となる。

$$\begin{aligned} Q^{\pi}(s_t,a_t) &\leftarrow (1-\alpha)Q^{\pi}(s_t,a_t) + \alpha(r_{t+1} + \gamma Q^{\pi}(s_{t+1},a_{t+1})) \\ &= Q^{\pi}(s_t,a_t) + \alpha(r_{t+1} + \gamma Q^{\pi}(s_{t+1},a_{t+1}) - Q^{\pi}(s_t,a_t)) \end{aligned} \tag{2.24}$$

(2.24)式は$(s_t,a_t,r_{t+1},s_{t+1},a_{t+1})$ で構成されているため、「Sarsa 学習」と呼ばれる[15][31]。Sarsa 学習のアルゴリズムは付録 A に示す。

方策 π によって次状態 s_{t+1} に対応する行動 a_{t+1} を選択し、その状態－行動価値関数 $Q^{\pi}(s_{t+1},a_{t+1})$ を用いる「方策オン」の Sarsa 学習に対し、次状態の取りうる行動のち、方策を用いず、最大の $\max_{b\in A} Q(s_{t+1},b)$ を用いて、現在の状態―行動価値関数を更新する TD 学習は

「Q 学習（Q-Learning）」と呼ばれる[15][28]。

$$Q^{\pi}(s_t,a_t) \leftarrow Q^{\pi}(s_t,a_t) + \alpha(r_{t+1} + \gamma \max_{b \in A} Q(s_{t+1},b) - Q^{\pi}(s_t,a_t)) \tag{2.25}$$

Q 学習のアルゴリズムは付録 B に示す。

2.4.8 価値関数の近似

前述した状態―行動価値関数 Q(s, a)の値は、離散状態空間 S と離散行動空間 A に対応した Q テーブルで表せる（表 2.1 参照）。

表 2.1 状態―行動価値関数の値 Q-table

状態／行動集合	a_1	a_2	・・・	$\|A\|$
s_1	$Q(s_1,a_1)$	$Q(s_1,a_2)$	・・・	$Q(s_1,\|A\|)$
s_2	$Q(s_2,a_1)$	$Q(s_2,a_2)$	・・・	$Q(s_2,\|A\|)$
・・・	・・・	・・・	・・・	・・・
$\|S\|$	$Q(\|S\|,a_1)$	$Q(\|S\|,a_2)$	・・・	$Q(\|S\|,\|A\|)$

連続空間や、部分観測マルコフ決定過程(POMDP: Partially Observable Markov Decision Process)に対応するため、状態、または状態価値関数を近似関数（モデル）で構成することが提案されている[15] [34] [35]。

【状態―行動価値関数の線形近似モデル】

状態と行動によって構成されるベクトル $\mathbf{x}=(s,a)^{\mathrm{T}} \in S \times A$ の特徴が $\boldsymbol{\varphi}(\mathbf{x}) = (\phi_1(s,a), \phi_2(s,a), .. \phi_{|S|\times|A|}(s,a))^{\mathbf{T}}$ とし、状態―行動価値関数 $Q^{\pi}(s,a)$ は以下のように近似される。

$$\hat{Q}^{\pi}(s,a;\boldsymbol{\theta}) = \sum_{b=1}^{|S|\times|A|} \theta_b \phi_b(s,a) = \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{\varphi}(\mathbf{x}) \tag{2.26}$$

ここで、$\boldsymbol{\theta} = (\theta_1, \theta_2, .. \theta_{|S|\times|A|})^{\mathbf{T}}$ は線形モデルのパラメータである。特徴ベクトルは線形独立な基底関数として、ガウス関数などを用いることができる。

$$\boldsymbol{\varphi}(\mathbf{x}) = \mathbf{exp}(-\frac{||\mathbf{x}-\mathbf{c}||^2}{2\sigma^2}) \tag{2.27}$$

ここうで、$\mathbf{c} = (c_1, c_2, .. c_{|S|\times|A|})^{\mathbf{T}}$ と σ は、それぞれ、ガウス関数の中心と標準偏差である。

(2.13)式の右辺の第 1 項は, 任意の現状態 s から次状態 s'へ遷移するに伴う報酬 r の期待値であり、以下の期待報酬関数 $R(s,a)$ で表す。

$$\begin{aligned} r = R(s,a) &\equiv \mathop{\mathbf{E}}_{Pr(s'|s,a)} \{R(s,a;s')\} \\ &= Q^{\pi}(s,a) - \gamma \mathop{\mathbf{E}}_{Pr(s'|s,a)} \mathop{\mathbf{E}}_{\pi(a'|s')} \{Q^{\pi}(s',a')\} \end{aligned} \tag{2.28}$$

近似関数を用いる場合、近似報酬関数 $\hat{R}(s,a)$ は以下となる。

$$\hat{R}(s,a) \equiv \hat{Q}^{\pi}(s,a) - \gamma \mathop{\mathbf{E}}_{Pr(s'|s,a)} \mathop{\mathbf{E}}_{\pi(a'|s')} \{Q^{\pi}(s',a')\} \tag{2.29}$$

近似誤差 $g_{TD} = \frac{1}{2}(r - \hat{R}(s,a))^2$ を最小にするよう、勾配法によって、線形モデルのパ

ラメータをエピソードごとに更新する。

$$\begin{aligned}\boldsymbol{\theta} &\leftarrow \boldsymbol{\theta} - \alpha \frac{\partial\{\frac{1}{2}(r-\hat{R}(s,a))^2\}}{\partial\boldsymbol{\theta}} \\ &= \boldsymbol{\theta} + \alpha(r-\hat{R}(s,a))\frac{\partial\hat{R}(s,a)}{\partial\boldsymbol{\theta}} \\ &= \boldsymbol{\theta} + \alpha(r-\hat{Q}(s,a)+\gamma \mathop{\mathbf{E}}_{\Pr(s'|s,a)} \mathop{\mathbf{E}}_{\pi(a'|s')}\{Q^{\pi}(s',a')\})\boldsymbol{\varphi}(\mathbf{x}) \\ &= \boldsymbol{\theta} + \alpha(r-\hat{Q}(s,a)+\gamma Q^{\pi}(s',a')\boldsymbol{\varphi}(\mathbf{x})\end{aligned} \tag{2.30}$$

(2.30)式の右辺の第 2 項に注意すれば、線形近似モデルも状態ー行動価値関数の TD 誤差 $\varepsilon^{TD} = r + \gamma Q(s',a') - Q(s,a)$ を用いて学習できることが分かる。

(2.26)式は従来の状態―行動価値関数 Q の線形近似[35]で、すべての状態と行動が既知と前提している。本論文では、環境状態が未知または変化する場合に対応するため、自己組織化ファジィニューラルネットワークを用いて Q 値を近似することを提案する（第 3 章参照）。

2.4.9 マルチエージェント強化学習

大規模かつ複雑で、常に変化するシステムに対しては、中央集権型システムによる最適制御や故障の診断などの対応は困難である。例えばサッカーゲームにおける各プレイヤーの行動は、ボールや相手などの環境要素の不確定さによって一概に規定できない。独立したエージェントやシステムがそれぞれ、部分問題を解決し、相互作用によって知的な集団行動を生み出すことによって、大規模システム問題を解決するアプローチは、1980 年代から始まり、これまで、多くの概念と研究成果が挙げられている。例えば、「分散制御型システム(decentralized system)」、「分散人工知能（DAI: Distributed Artificial Intelligence）」、「マルチエージェントシステム」、「群知能(Swarm Intelligence)」、「スワームロボティクス(SR: Swarm Robotics)」などが良く知られている[4]-[7] [10]-[13] [82]-[92] [110]-[118] [128]-[131]。

一方、自律エージェントや自律システムが複数存在する際に、「観測状態数の爆発」や「次元の呪い(The curse of dimensionality)」、「状態遷移の不確定性(uncertainty of state transition)」及び「不完全知覚問題(perceptual aliasing problem)」（図 2.4 参照）などの問題が生じる。これらの問題をいかに解決できるかはマルチエージェント系学習の挑戦的な課題である。

本論文では以下の技術を用いて、知的個体集団の学習方式を提案し、知的個体群による未知環境の目標探索問題の解決を試みる。

(i)　状態の曖昧さに対応する「ファジィ推論」;

(ii) 任意の入出力関係を定める「ニューラルネットワーク」;

(iii) 試行錯誤及び報酬情報によって適応行動を獲得する「強化学習」;

(iv) 群れの形成及び保持原理「BOID ルール」。

なお、「BOID」とは、Reynolds が提案した鳥の群れのシミュレーション「birdoid」の略称

であり、個体の移動が以下の三つのルールに従えば、群れの形成及び動的保持が実現できるという[110] [111]。

(i) 衝突回避：近すぎる群れの仲間と離れ、衝突を回避する;

(ii) 速度調整: 周りと速度を合わせる;

(iii) 求心力：群れの中心の方へ向かおうとする。

このような単純なルールによって各個体が自分の方向を分散的に決めるように設定しておくと、現実の鳥と同じような群れが観察される。さらに、単に群れているというだけでなく、障害物を避けるために群れが 2 つに分かれてまた合流するという動きも見られ、現実の鳥の群れを想起させる。

本論文では、これらのルールを更に簡略化し、「離れすぎず、近すぎず」という適切な距離を保つ行動に正の報酬を与え、そうでない場合の行動に負の報酬を与えるのみで、複数の知的個体の群れの形成及びその群れの保持を実現させることを試みる。

2.4.10 強化学習の応用分野

従来の学習システムの理論[60]と比べ、強化学習は動物の行動分析から提案された知的計算理論であり、これまで、特に適応制御[61]-[66]、自律ロボット[10] [37] [43]-[54] [67]-[81]、群知能[82]-[94]、非線形予測[38]-[42] [93]、Web サービス[96] [97]など多くの研究分野で応用されている。

第 3 章

SOFNN を用いた強化学習システム

未知環境における目標の探索問題を扱う知的個体（図 2.1 参照）の内部モデルの「状態認知モジュール」は、ファジィ推論システムやニューラルネットワークなどの技術によって構築する際に、ファジィメンバーシップ関数やファジィルールの設定、教師データの収集、ネットワークの構造の決定など多くの課題がある。本章では、第 2 章で述べたデータ駆動による自己組織化ファジィニューラルネットワーク(SOFNN)を知的個体の状態認知モジュールとし、適応行動を出力する強化学習システムと融合することにより、知的個体の内部モデルを構成することを提案する。

第 3.1 節では、まず、SOFNN と Actor-Critic 型強化学習方式を融合し、状態観測から、適応行動を決定するまでの知的個体の内部モデル FAC を構成する。次に、2.4.7 節に述べられた TD 学習を FAC に導入し、方策関数の学習を行う。そして、完全観測環境及び部分観測環境における目標探索問題のシミュレーションを行い、FAC の性能を確認する。更に、複数の知的個体の場合のシミュレーションを行い、群れの形成及び群学習の効果を明らかにする。本節の主な内容は文献[85]-[88] [94]で発表されている。

第 3.2 節では、Actor-Critic 型強化学習の代わりに、TD 学習の一つである Q 学習（QL: Q-learning）を用いて、新たな強化学習システム FQ を提案する。部分観測マルコフ決定過程(POMDP)下の未知環境探索問題のシミュレーションを行い、FQ の性能を確認する。更に、複数の知的個体の場合のシミュレーションを行い、群れの形成及び群学習の効果を明らかにする。本節の主な内容は文献[92]-[94]で発表されている。

第 3.3 節では、よく知られるもう一つの強化学習方式である Sarsa 学習[15]を用いて、新たなニューロファジィ型強化学習システム FS を提案する。部分観測マルコフ決定過程(POMDP)下の未知環境探索問題のシミュレーションを行い、FS の性能を確認する。また、複数の知的個体の場合を考慮し、群れの形成及び群学習の効果を明らかにする。本節の主な内容は文献[92]～[94]で発表されている。

3.1　FNN を用いた Actor-Critic 型強化学習システム(FAC)

3.1.1 FAC の構成

提案するファジィネット(FN: Fuzzy net)を用いた Actor-Critic 型強化学習システム(FAC)の構成及び FAC が知的個体（Agent）として環境と相互作用する様子を図 3.1 に示す。ここで、「Fuzzy net」とは前章で述べた自己組織化ファジィニューラルネットワーク(SOFNN)の略称であり、「Actor」と「Critic」は FN の出力（推論結果）に重み付けした「行動価値関数」と「状態価値関数」である。FAC の詳細な構成を図 3.2 に示す。

まず環境から観測した状態が入力情報として Fuzzy net 部に入力される。Fuzzy net 部では入力情報とメンバーシップ関数からルールの適合度 $\phi_k(x_t)$ を計算し、推論結果に結合荷重(図 3.2 の結合線上の黒い点)を付け、強化学習の Actor の $A_m(x_t)$ と Critic の $V(x_t)$ へ出力する。エージェント（知的個体）は確率方策関数 $\pi(A_m(x_t))$ の確率分布 $p(a_t = a_j \mid x_t) = f(A_m(x_t))$ に従って確率的に行動 a_j を選択する。

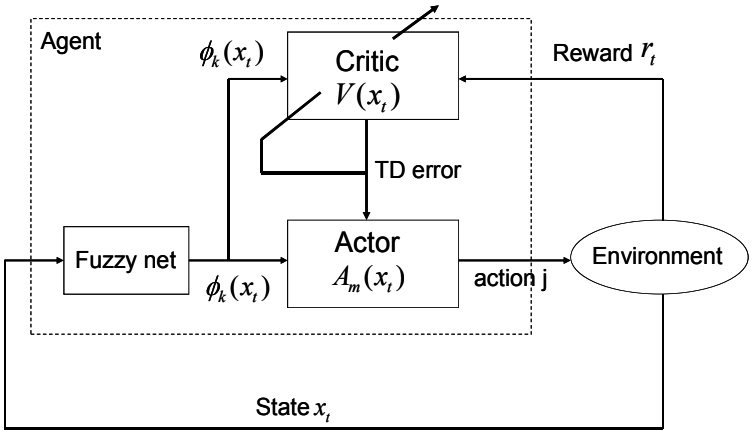


図 3.1 FAC を用いる知的個体(Agent)

行動結果から Critic は次状態の価値関数 $V(x_{t+1})$ を推定し、$V(x_t)$ と $V(x_{t+1})$ から状態遷移において連続する状態間の関数値の時間的差分となる TD（Temporal Difference）誤差を計算する。TD 誤差から、Actor はより高い報酬を得る行動を選択するように学習し、Critic は TD 誤差が 0 に近づくように結合荷重の修正、すなわち、学習を行う。

まず、ファジィ推論の出力である φ(x)に対し、状態価値関数 $V(x_t)$ への結合荷重を $v_k, k = 1, 2, ..., K$ とし、行動価値関数 $A_m(x_t)$ への結合荷重を $w_{kj}, k = 1, 2, ..., K, j = 1, 2, ..., J$ とする。ただし、K と J はそれぞれルール数、行動数を意味する。

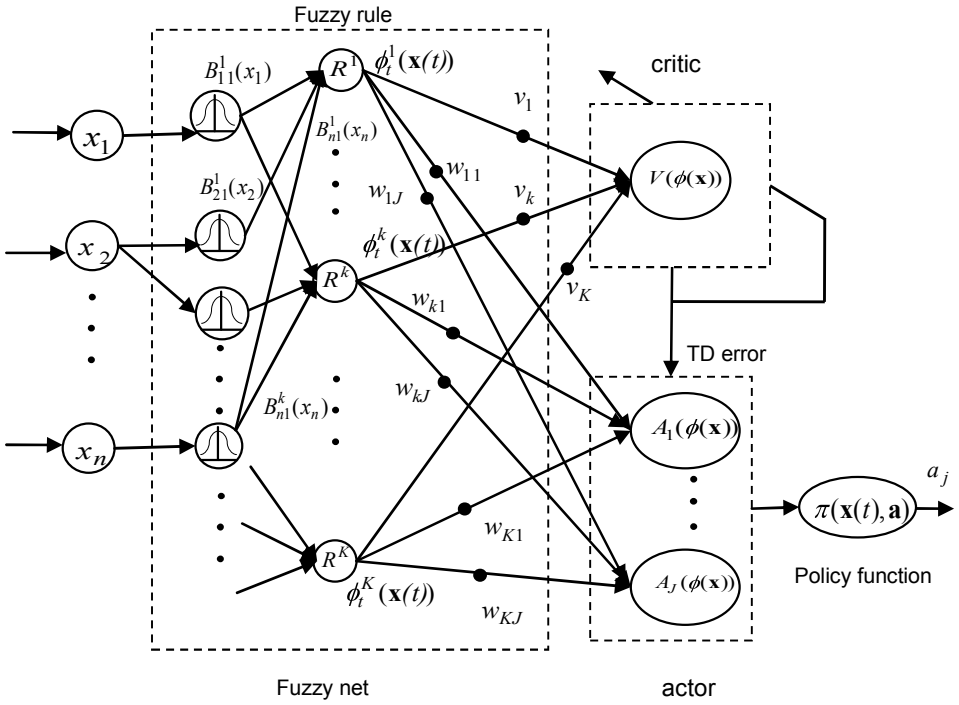


図 3.2　FAC の構成

Critic と Actor の出力は、以下の式で計算される。

$$V(\mathrm{x}(t)) = \frac{\sum_k v_k \phi_t^k(\mathrm{x}(t))}{\sum_k \phi_t^k(\mathrm{x}(t))}, \tag{3.1}$$

$$A_j(\mathbf{x}(t)) = \frac{\sum_k w_k \phi_t^k(\mathbf{x}(t))}{\sum_k \phi_t^k(\mathbf{x}(t))}. \tag{3.2}$$

次に、確率方策関数 $\pi(A_m(x_t))$はボルツマン分布を用いる。すなわち、状態 $\phi(\mathrm{x})$ に対し、行動 $a_j, j = 1, 2, ..., J$ が選択される確率は以下の式になる。

$$p(a_t = a_j \mid \mathrm{x}(t)) = \frac{\exp(A_j(\mathrm{x}(t))/T)}{\sum_b \exp(A_b(\mathrm{x}(t))/T)}. \tag{3.3}$$

但し、ここで $T>0$ は温度定数と呼ばれるパラメータである。

そして、高い報酬を獲得する行動を高い確率で選択するため、TD 誤差を用いた学習則によって結合荷重 v_k と w_{kj} を修正する。FAC の学習則については、次節で述べる。

3.1.2 FAC の学習則

任意の状態から、行動によって、エージェントは次状態に遷移する。未知環境での目標探索過程は有限状態の遷移過程に当たる。一般的に、2 章で述べた強化学習の TD 学習は行動方策を改善するため、 (2.20)式に定義された時刻 t における状態 s_t の価値 $V^{\pi}(s_t)$ を、(2.21)式に定義された TD 誤差によって、(2.22)式のように直接更新するが、本論文で提案する FAC 型強化学習システムにおいて、ファジィネットの出力に対する結合荷重を更新することによって、状態価値関数を間接的に更新する。すなわち、FAC の学習則は以下の(3.4)-(3.6)式で定める。

$$v_k^{new} \leftarrow v_k^{old} + \beta_v \varepsilon_{TD}(\mathbf{x}(t))\phi_t^k(\mathbf{x}(t)) \tag{3.4}$$

$$w_{kj}^{new} \leftarrow w_{kj}^{old} + \begin{cases} \beta_w \varepsilon_{TD}(\mathbf{x}(t))\phi_t^k(\mathbf{x}(t)) & a_t = a_j \\ 0 & otherwise \end{cases} \tag{3.5}$$

但し、 $0 < \beta_v, \beta_w \le 1$ は学習率であり、TD 誤差 $\varepsilon_{TD}(\mathrm{x}(t))$ は以下のように定義される。

$$\varepsilon_{TD}(\mathbf{x}(t)) = r_t + \gamma V(\mathbf{x}(t+1)) - V(\mathbf{x}(t)) \tag{3.6}$$

なお、 r_t は状態 $\mathrm{x}(t)$ で行動 a_t によって状態 $\mathrm{x}(t+1)$ に到達する際に獲得した報酬で、γ は減衰率である。

3.1.3 FAC の群学習と単独学習

個体間の距離を適切に保つような行動に正の報酬を与え、そうでない場合は負の報酬を与えることによって、複数個体による多点探索を実現可能とし、探索の効率を高めることを考える。ここで、個体間の距離を考慮する場合の学習則を「群学習(Swarm Learning)」と呼び、そうでない場合は「単独学習(Individual Learning)」と呼ぶことにする。

個体間の距離を測るため、2 次元環境における座標情報(X,Y)を用いることにする。先ず複数個体の群重心($\overline{X},\overline{Y}$)を求め、群重心と座標(X_a,Y_a)に存在する個体 a との距離 D_a を次式で計算する。

$$D_a = \sqrt{(X_a - \overline{X})^2 + (Y_a - \overline{Y})^2} \tag{3.7}$$

ここで、 $\overline{X} = \sum_{a=1}^{N} X_a / N, \overline{Y} = \sum_{a=1}^{N} Y_a / N$ 、N はエージェント数である。

TD 誤差を求める(3.6)式における報酬 r_t は、「離れ過ぎず近過ぎず」という BOID の群れの形成ルール[(1)]を満たすため、次式で与える。

$$r_t = r_{swarm} + r_{noswarm} + r_{goal} + r_{goal}^{G} + r_{crash}\,. \tag{3.8}$$

ここで、 r_{swarm} は $D_{\min} \le D_a \le D_{\max}$ の場合の正の報酬、$D_{\max}$ と $D_{\min}$ は個体 a の重心との最大距離と最小距離の閾値であり、 $r_{noswarm}$ は群れと離れる行動をとった場合（ $D_a < D_{\min}$ または $D_a > D_{\max}$ ）の負の報酬（例えば $r_{noswarm} = -D_a$ とする）、 r_{goal} と r_{goal}^{G} は目標エリアと目標エリア内にある目標 G に到達したときの正の報酬値（ $r_{goal}^{G} > r_{goal} >> r_{swarm} > 0$ ）、r_{crash} は障害物や他の個体と衝突する場合の負の報酬である。

3.1.4　FAC の計算機シミュレーション

本章で述べた自己組織化ファジィニューラルネットワークを用いた Actor-Critic 型強化学

習システム(FAC)の性能を確認するため、2 次元平面空間における複数知的個体の目標探索シミュレーションを行う。なお、環境の状態は、座標情報を観測するマルコフ決定過程(MDP)の場合と、学習者の周辺のみ、近傍情報を観測する部分観測マルコフ決定過程(POMDP)の場合に分類され（2.4.2 節）、それぞれの場合のシミュレーション及び評価結果を示す。

(i)マルコフ決定過程（MDP）：離散空間の場合

2 体の個体がそれぞれ点(1, 4)と点(3,1)から出発し、36x36 平面上の目標エリア(31,31)×(36,36)を探索し、その最短経路を決定する問題とする（図 3.3）。個体がこの未知の環境を探索する際に、自己及び他個体の位置情報を 2D 座標によって定め、行動としては上下左右 4 方向の 1 ステップ 1 マスの移動、すなわち 4 つの選択可能な行動とし、2 体が共にエリア内に到達する場合は探索が終了する。なお、出発点から目標エリアに到達し、探索が終了するこの過程を「試行(cycle)」と呼ぶ。強化学習はこの過程を反復することによって最短経路を求める。

なお、最短経路長は複数個体の間の適切な距離を考慮した場合、最も遠い個体が出発点の(3,1)からゴールエリアの最近点（31,31）から 2 マス以上離れた更に奥の場所(33,33)に到達することを想定し、33-3+33-1=62 ステップのマンハッタン距離になる。

シミュレーションに用いたパラメータは表 3.1 に示す。目標エリアに到達する場合は報酬 100、壁に衝突する際には-1、1 体が目標エリアに進入し、もう 1 体が未到達の場合、進入した個体が目標エリアから出る場合の報酬は-1 とする。なお、個体間のユークリッド距離が(1.5～3.0)であれば、群れの形成に適切として、報酬 1 とし、その距離を保たない場合は報酬-1 とする。

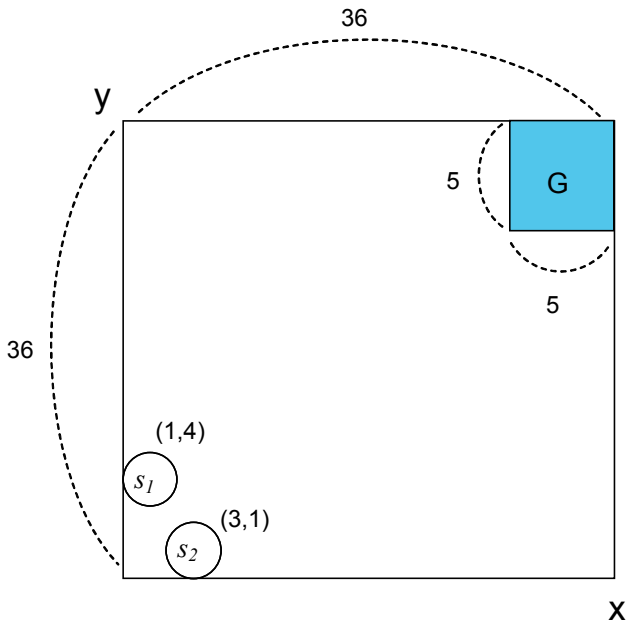
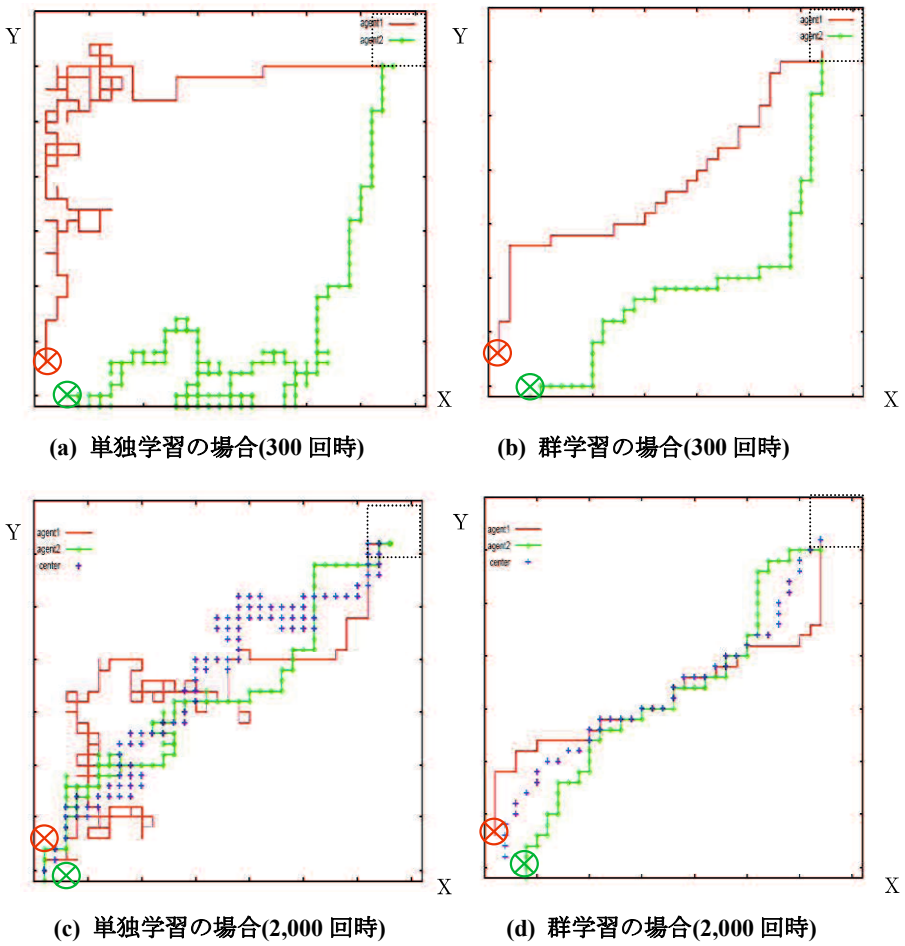


図 3.3 FAC を用いた目標探索シミュレーション(離散空間の場合)



(a) 単独学習の場合(300 回時)　(b) 群学習の場合(300 回時)

(c) 単独学習の場合(2,000 回時)　(d) 群学習の場合(2,000 回時)

図 3.4 FAC を用いた 2 つ個体の探索結果（離散空間の場合）

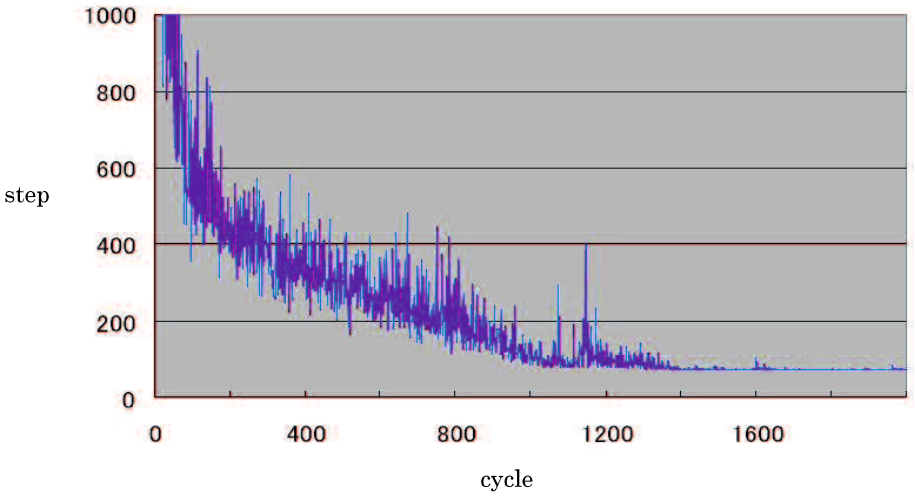
FAC を用いた 2 体の個体の探索結果を図 3.4(スタート位置は⊗、ゴールエリアは点線枠内)に示す。図 3.4(a)と図 3.4(b)は、学習途中、300 回試行の結果で、適切距離を考慮しない場合（「単独学習」）と、適切距離を考慮する場合（「群学習」）を示している。学習終了時(2,000 回試行)の目標探索結果（軌跡）を図 3.4(c)と図 3.4(b)に示している。なお、図 3.4(c)と(d)にある”+”は、両個体の中心位置を現し、時間的に個体間の距離の変化をより分かりやすく示している。図 3.4 のいずれの場合も適切距離を考慮した「群学習」の方は、経路長が短く、探索の効率が高くなっていることが分かる。

表 3. 1 離散状態―行動空間の場合 FAC のパラメータ

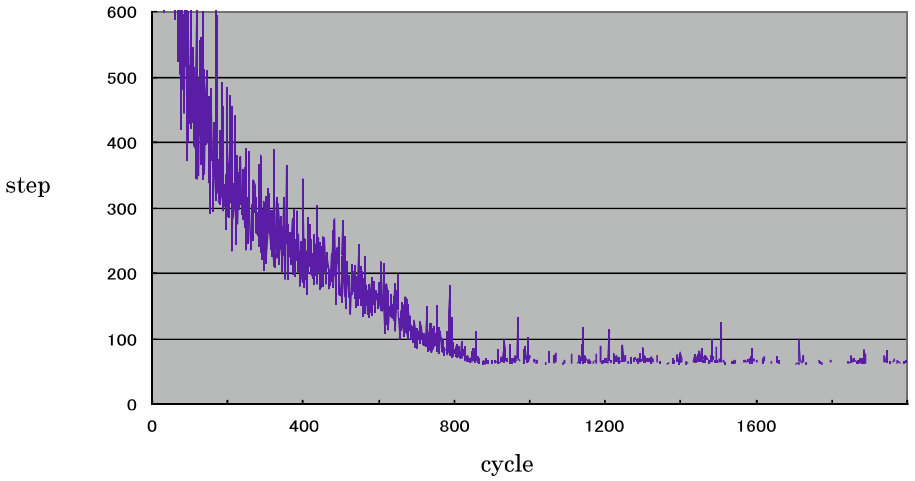
記　述	符号	値
入力ベクトルの次元数	n	2
出力行動空間の次元数	J	4
メンバーシップ関数の広がり	σ^2	0.1
メンバーシップ関数の増殖閾値	F	0.4
Critic への結合荷重の初期値	v_k	1.0
Actor への結合荷重の初期値	w_{kj}	0.25
Critic の学習率	β_v	0.3
Actor の学習率	β_w	0.3
TD-error の減衰率	γ	0.9
確率方策における温度定数	T	0.1
ゴールエリアに到達する報酬	r_{goal}	100.0
障害物に衝突する報酬	$r_{obstacle}$	-1.0
適切距離を保つ報酬	r_{swarm}	1.0
適切距離を保たない報酬	$r_{no\text{-}swarm}$	-1.0
最小適切距離閾値	min_dis	1.5
最大適切距離閾値	max_dis	3.0

FAC の学習性能(10 回シミュレーションの平均探索ステップ数)を図 3.5 に示す。図 3.5(a)の「1 体で学習」の場合は、学習収束が 1,400 回試行から実現していることに対して、図 3.5(b)の「2 体で単独学習」の場合と図 3.5(c)の「2 体で群学習」の場合は 850 回試行からより早く収束が見られる。また、収束時の平均ステップ数（探索経路長）は単独学習と群学習の場合それぞれ 63.6 と 63.3 であった。

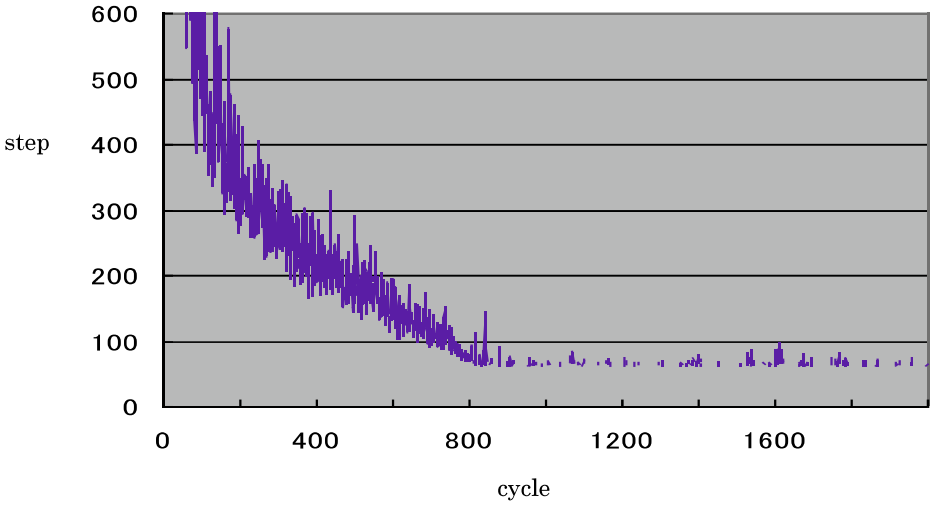
なお、本シミュレーションにおいて、Fuzzy net の自己組織化の処理はメンバーシップ関数とファジィルールの自己増殖を適切な閾値の設定によって実現した。そのため、冗長と思われるメンバーシップ関数とファジィルールの統合と削除の処理はなかった。



(a) 1 体で学習の場合

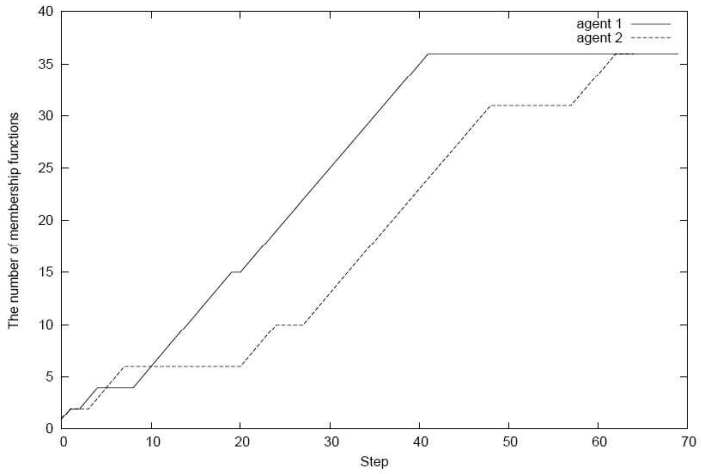


(b) 2 体で単独学習の場合

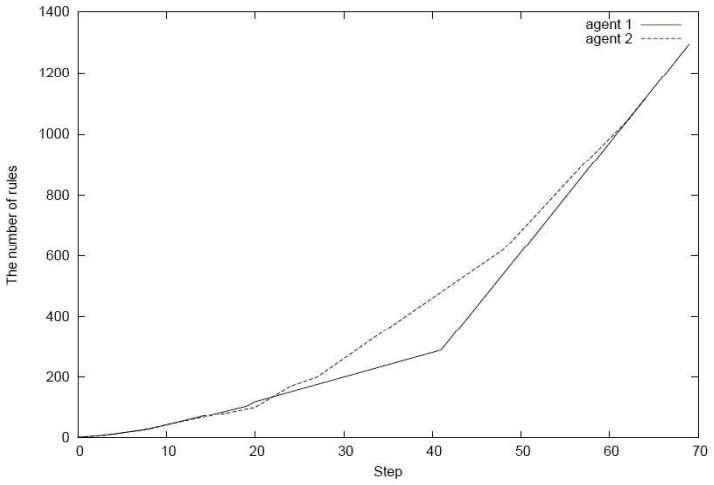


(c) 2 体で群学習の場合

図 3.5 FAC の学習性能（離散空間の場合）



(a) x 軸入力に対するメンバーシップ関数数の増加



(b) ファジィルール数の増加

図 3.6 FAC のメンバーシップ関数とファジィルールの増殖（離散空間の場合）

図 3.6 はシミュレーションにおいて、FAC の離散環境の状態を観測し、ファジィメンバーシップ関数とファジィルールが増加し続ける様子を示す。入力次元 x 軸と y 軸のメンバーシップ関数の数（図 3.6(a)は x 軸の場合を示す）は共に環境サイズ 36x36 を反映した 36 個に増加し、ファジィルールの数は 36x36=1,296 個となっている。

また、FAC のロバスト性について、「学習終了後、任意の出発点から、両個体がゴールエリアへの行動、及び経路を観測する」ことによって、学習効果が確認できた。図 3.7 は収束の早い群学習を用いた場合、両エージェントの出発点は学習時の(1,4)、(3,1)から、学習終了後の(1,22)と（19,1）に変更した場合の軌跡を示す。初期の行動はランダム探索に近いが、途中から両個体が寄り添ってゴールエリアに向かったことができた。

なお、個体数が増加した場合のシミュレーションも行い、FAC の有効性を確認した。図

3.8 は 4 体の場合の群学習を行った後の目標探索軌跡を示している。個体の出発位置は(34, 34)、(34, 32)、(33, 33)、(35, 33)で、ゴールエリアは(0, 0)～(5, 5)の正方形領域とした。

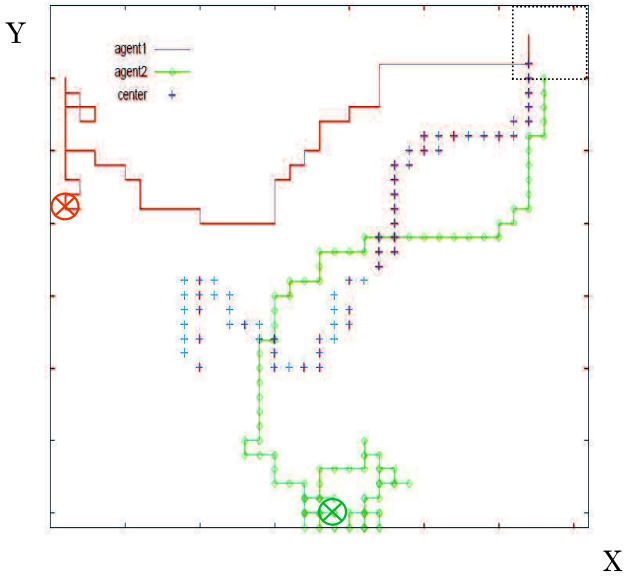


図 3.7 FAC の頑健性について（離散空間の場合）

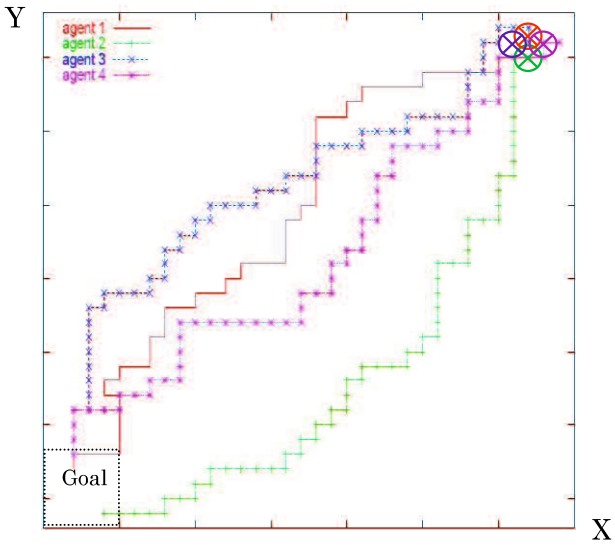


図 3.8 FAC を用いた群学習の終了時の探索軌跡（離散空間の場合）

(ii) マルコフ決定過程（MDP）：連続空間の場合

観測状態が座標の連続値で表し、行動が任意方向となる連続状態―行動空間における FAC 個体の目標探索シミュレーションを行った。図 3.9 は障害物の存在する探索環境を示している。２つ個体の出発点はそれぞれ(1,1)と(2,2)とし、ゴールエリアは(35,35)から(36,36)の間とした。

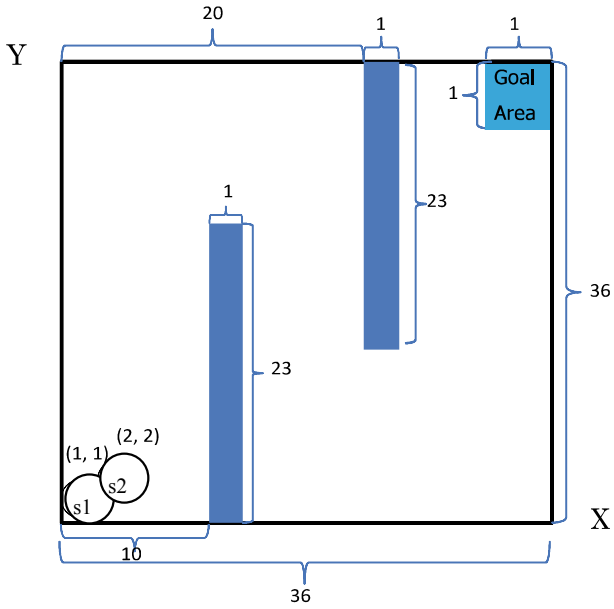


図 3.9 FAC を用いた目標探索シミュレーション(連続空間の場合)

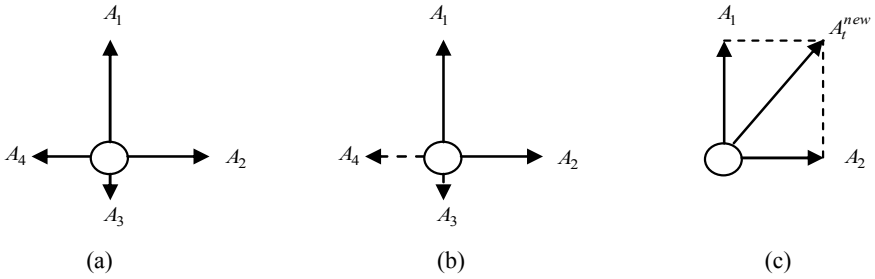


図 3.10 行動価値関数を用いた移動方向の決定(連続空間の場合)

FAC の出力である行動は図 3.10 に示すように、上下左右の 4 方向の行動価値（図 3.10(a)）を用いて、高い価値順の 2 つの行動方向を選び（図 3.10(b)）、それらの線形結合によって個体の移動方向を決めた（図 3.10(c)）。また、 1 step ごとの移動距離は常に 1 とした。連続空間を扱う FAC のパラメータの設定は表 3.2 に示す。

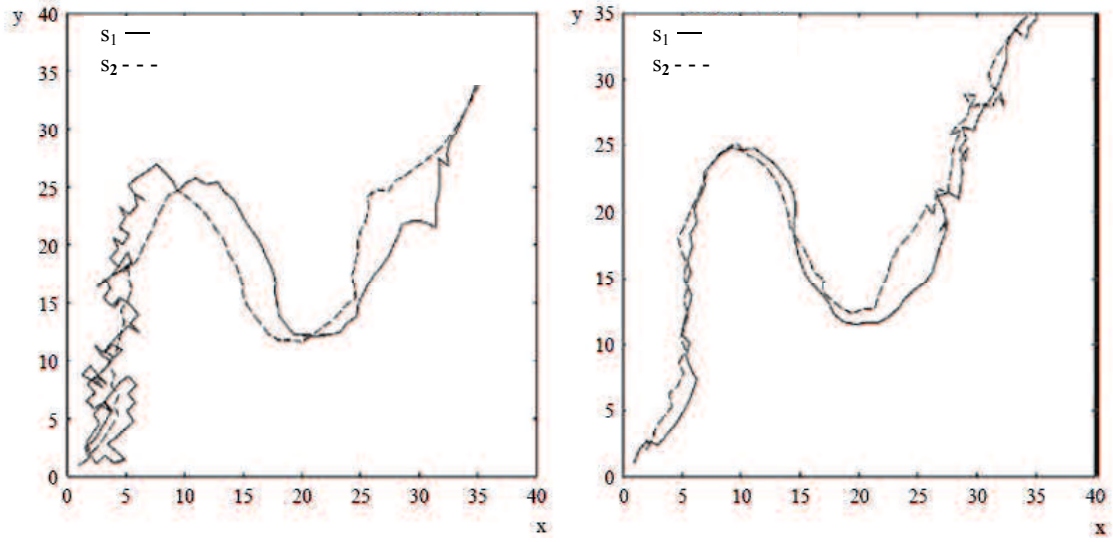
障害物を回避し、ゴールエリアへの経路を探索した結果を図 3.11 に示す。300 回の学習回数と共に、経路長が減少及び収束する様子を図 3.12 に示す。個体間に適切な距離を保つことで報酬を与える「群学習」の FAC（図 3.11(b)、図 3.12「swarm.dat」）と、そうでない「単独学習」の FAC（図 3.11(a)、図 3.12「indi.dat」）との学習性能は、前者の方が比較的高いことが、図 3.11 と図 3.12 で確認できる。

なお、前節の離散値座標を用いる場合の FAC のファジィメンバーシップ関数の数（x 軸 36 個、y 軸 36 個）とファジィルール数（36x36=1,296）より、連続値座標を用いた FAC の

メンバーシップ関数とファジィルールの数は、それぞれ 109x2 と 2,970 まで増加した。

表 3. 2 連続状態―行動空間の場合 FAC のパラメータ

記　述	符号	値
入力ベクトルの次元数	n	2
出力行動空間の次元数	J	4
メンバーシップ関数の広がり	σ^2	0.1
メンバーシップ関数の増殖閾値	F	0.4
Critic への結合荷重の初期値	v_k	1.0
Actor への結合荷重の初期値	w_{kj}	0.25
Critic の学習率	β_v	0.3
Actor の学習率	β_w	0.3
TD-error の減衰率	γ	0.9
確率方策における温度定数	T	0.1
ゴールエリアに到達する報酬	r_{goal}	100.0
障害物に衝突する報酬	$r_{obstacle}$	-10.0
角に衝突する報酬	r_{corner}	-20.0
適切距離を保つ報酬	r_{swarm}	1.0
適切距離を保たない報酬	$r_{no\text{-}swarm}$	$-D$(距離)
最小適切距離閾値	min_dis	1.5
最大適切距離閾値	max_dis	3.0



(a) 単独学習の結果　　**(b) 群学習の結果**

図 3.11 単独学習と群学習の比較(連続空間の場合)：探索軌跡

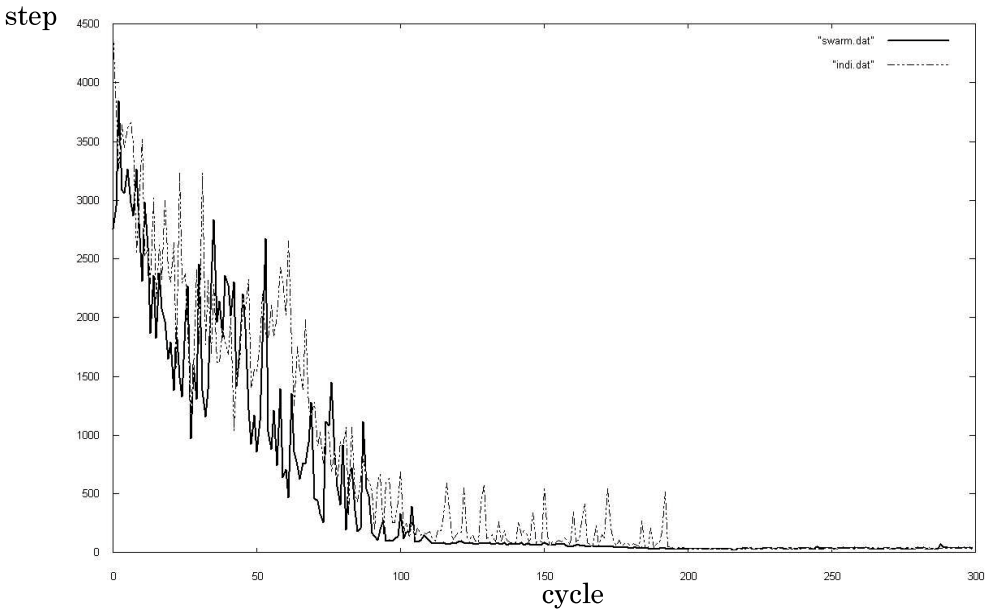
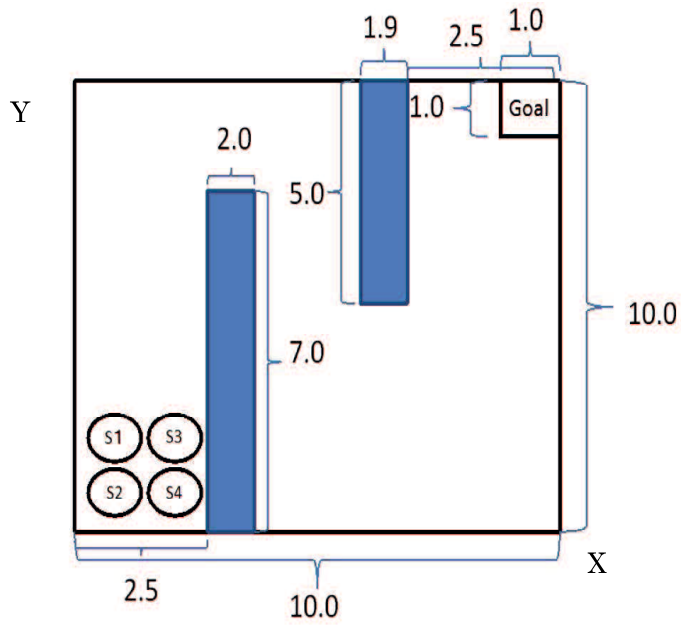
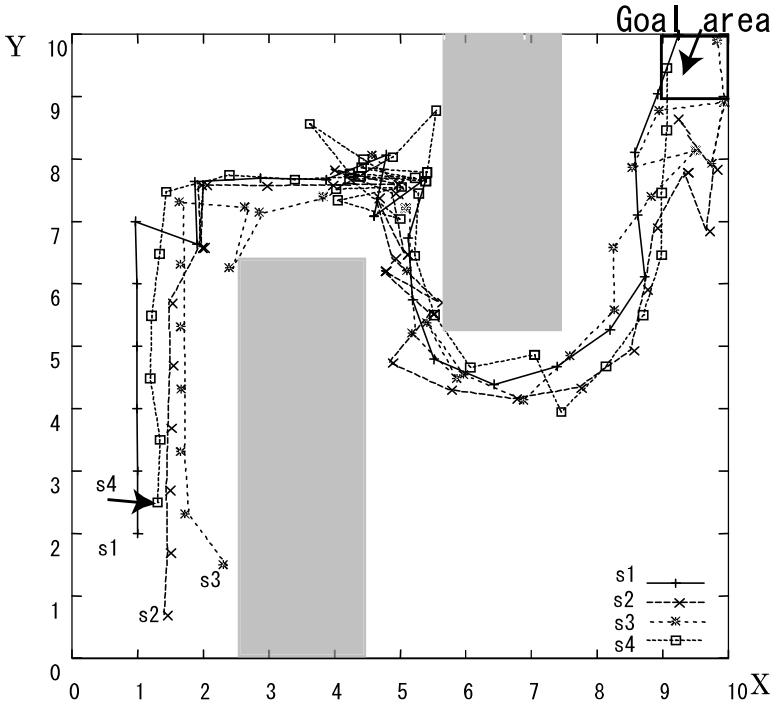


図 3.12 単独学習と群学習の学習性能の比較(連続空間の場合)

図 3.13 は個体数が 4 体の場合の連続値群行動の環境設定（図 3.13 (a)）及び学習結果（図 3.13 (b)）を示している。計算コストを減らすため、探索空間のサイズは 10x10 の比較的小さい障害物のある空間とした。左下(1, 2)、(1.4, 0.7)、(2.3, 1.5)、(1.3, 2.5)のそれぞれの場所から出発した 4 体の個体が適切な距離を保ちながらゴールエリアに到達できた。



(a)　4 体の個体の探索環境（連続座標値を用いる）



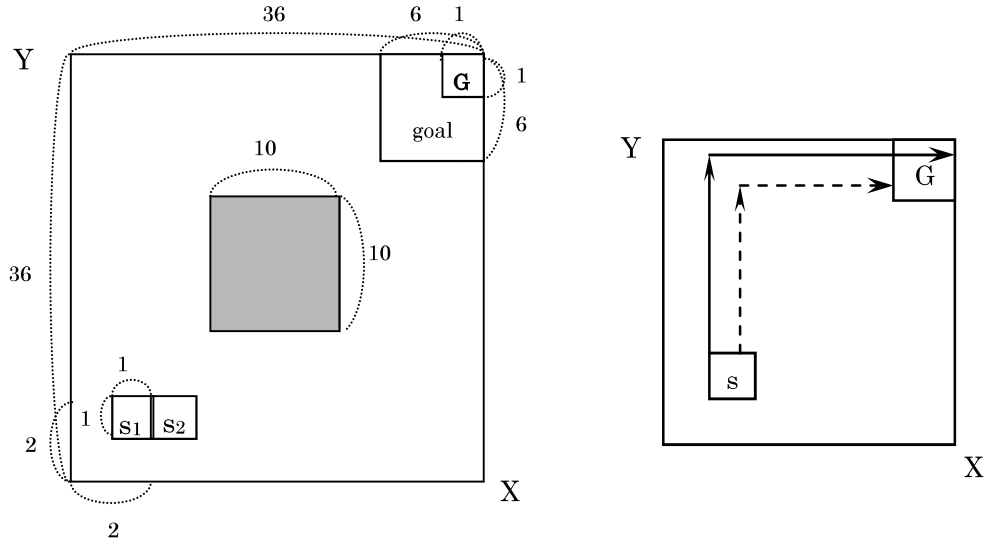
(b) 4 体の個体の群学習終了時の探索経路（連続空間の場合）

図 3.13　個体 4 体を用いたシミュレーション(連続空間の場合)

(iii)部分観測マルコフ決定過程（POMDP）の場合

不完全知覚によって、学習主体が観測した状態が部分的で、同じ状態とみなしても、実際は異なる状態に対し、同じ行動を出力してしまい、状態の遷移が実行される確率過程は

「部分観測マルコフ決定過程(POMDP: Partially Observable Markov Decision Process)」と呼ばれる（第 2 章 2.3.2 節、図 2.4 参照）。本章で述べた FAC を用いた POMDP 環境での目標探索シミュレーションを行い、FAC の有効性及び単独学習、群学習の性能を調べた。



(a) 問題設定　　**(b) 最短経路（破線）と準最短経路（実線）**

図 3.14 POMDP における未知環境探索問題のシミュレーション

図 3.14(a)は障害物が存在する探索環境を示し、2 体の個体は(2, 2)と(3, 2)の出発点から、中央に 10x10 の障害物を回避し、(30, 30)と(36, 36)の間のゴールエリアを到達する最短経路を探索する問題設定である。各個体の行動は上下左右方向で、距離は 1 ステップ 1 マスで、計 4 個ある。また、観測する状態は前節で述べた 2 次元ユークリッド空間の座標値（離散値または連続値）でなく、上下左右 4 方向の近傍情報であり、4 次元ベクトルである。近傍情報とは、個体の 1 マス離れた位置の環境は通路やゴールエリアであれば 0、壁や障害物の場合は 1 の値のことを指す。すなわち、各個体が観測できる状態数は 2^4=16 個あることになる。ゴールエリア内に「goal」と「G」で示される領域の報酬は異なる。その理由は、2 体の個体が個体間に一定の距離を保ちながら群れを形成してゴールエリアへ到達させるためである。

図 3.14(b)は 1 体の個体がゴールエリアに到達する最短経路（破線）と準最短経路（実線）の例を示す. 出発時の観測状態を(0, 0, 0, 0)とし、最短経路は、(0, 0, 0, 0)の状態が続く中で、個体が取る行動が「上」から「右」に変化し、同じ状態を観測したとしても異なる行動を取ることによって実現される。一方、個体が(0, 0, 0, 0)の状態遷移の途中、壁に沿う状態(1, 0, 0, 0)に変わってはじめて最適な行動が「上」から「右」に変化する場合、準最短路（実線）という準最適解しか得られない。ある一つの観測状態に対して、ある一つの適切な行動を出力する強化学習にとって、POMDP 下の最適解を求めることは困難であるが、準最適解へ

の接近は可能である。

表 3.3 に POMDP の場合の FAC のパラメータを示す。図 3.15 は 2 体の個体が FAC を用いた学習による探索経路長の減少及び収束状況を示す。単独学習(Individual Learning)場合と群学習（Swarm Learning）の場合共に最適解の 56step へは十分接近できなかったが、2500step 以降、ある程度の収束が見られ、群学習の解は単独学習と比べ、より最適解へ接近していることが分かる。学習終了時の両個体の探索経路を図 3.16 に示す。群学習（図 3.16(a)）に比べ、単独学習（図 3.16(b)）の経路は迂回しているため、経路長が 747.9step から 2,689.4step と長くなっている。

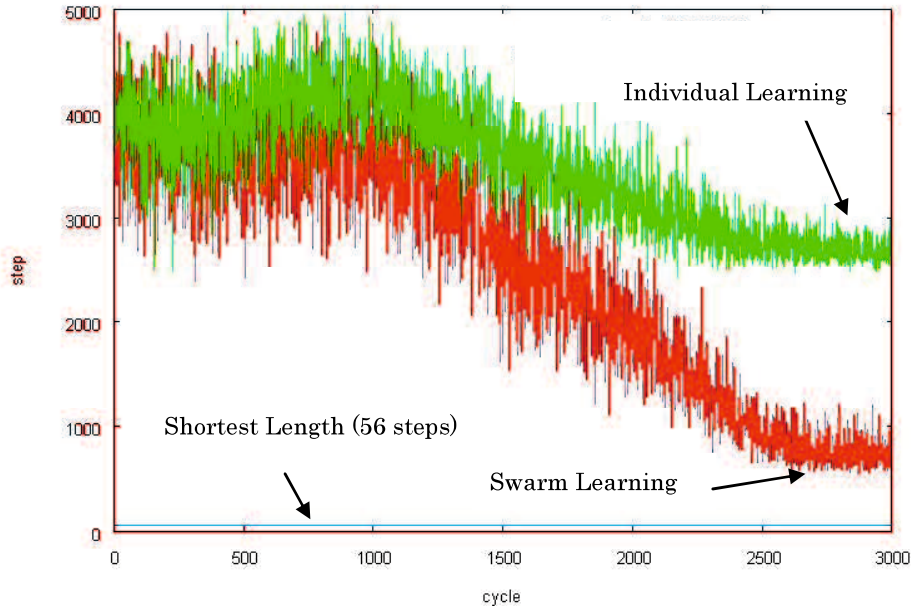
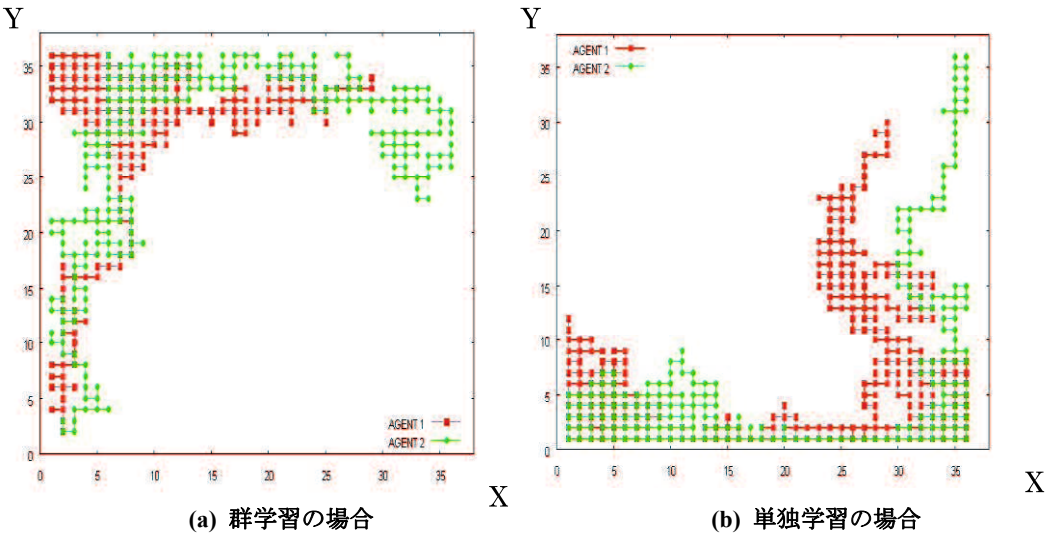


図 3.15 POMDP 下の FAC の学習性能

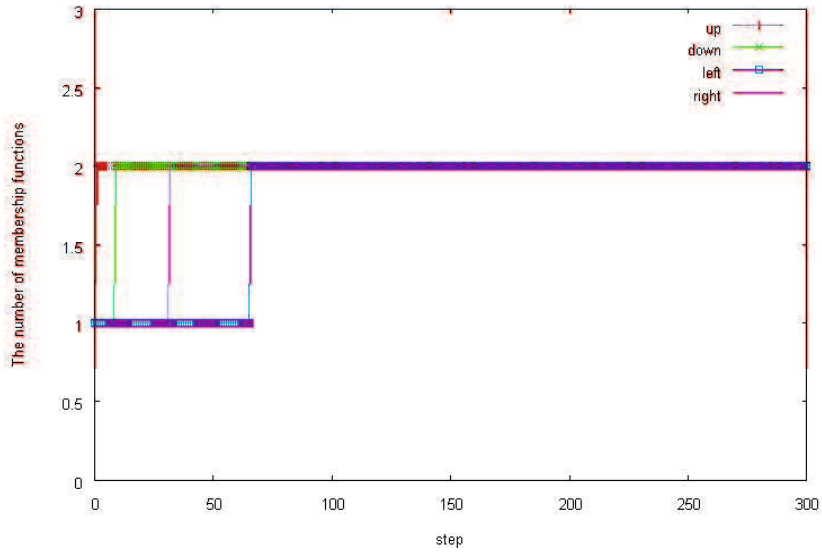


(a) 群学習の場合　　**(b) 単独学習の場合**

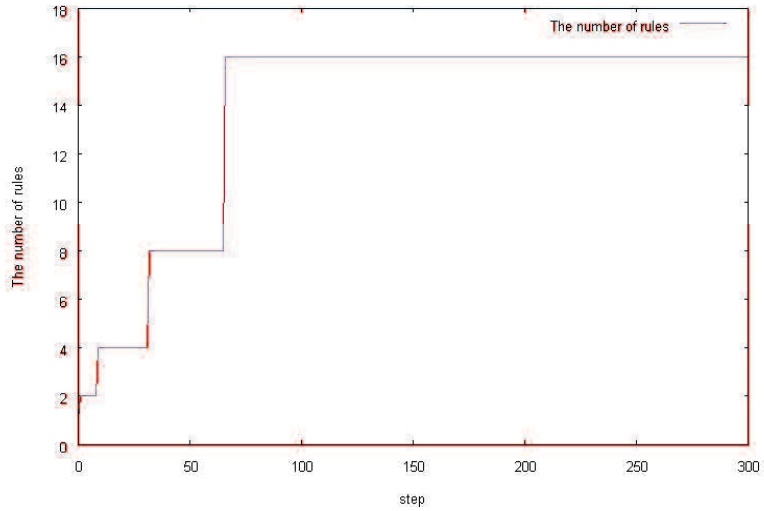
図 3.16 POMDP 下の FAC の学習結果

表 3. 3 POMDP の場合の FAC のパラメータ

記　述	符号	値
入力ベクトルの次元数	n	4
出力行動空間の次元数	J	4
メンバーシップ関数の広がり	σ^2	0.4
メンバーシップ関数の増殖閾値	F	0.4
Critic への結合荷重の初期値	v_k	1.0
Actor への結合荷重の初期値	w_{kj}	0.25
Critic の学習率	β_v	0.3→0.000001 (1000 試行目から)
Actor の学習率	β_w	0.3→0.000001 (1000 試行目から)
TD-error の減衰率	γ	0.99
確率方策における温度定数	T	0.7
ゴールエリア goal の報酬	r_{goal}	1000.0
ゴールエリア G の報酬	r_G	2000.0
障害物に衝突する報酬	$r_{obstacle}$	-10.0
適切距離を保つ報酬	r_{swarm}	1.0
適切距離を保たない報酬	$r_{no\text{-}swarm}$	-D (距離)
最小適切距離閾値	min_dis	1.5
最大適切距離閾値	max_dis	3.0



(a) メンバーシップ関数の生成



(b ファジィルールの生成

図 3.17 POMDP 環境の FAC の Fuzzy net

FAC への 4 次元入力に対し、Fuzzy net のメンバーシップ関数とファジィルールの増加の様子を図 3.17 に示す。いずれの次元とも入力値の 0 と 1 に対応する二つのメンバーシップ関数と、すべての状態を表す 16 個のファジィルールが生成されていることが分かる。

3.1.5 本節のまとめ

本節では前章で述べた自己組織化ファジィニューラルネットワーク（SOFNN）を用いて、Actor-Critic 型の強化学習方式と融合し、ファジィニューラルネットワーク型強化学習システム FAC を提案し、FAC を用いた複数個体の目標探索問題のシミュレーション及びそれらの結果を示した。目標探索問題の環境設定について、離散座標値と連続座標値を用いた状態が完全観測可能なマルコフ決定過程(MDP)下の環境と、近傍環境のみが観測可能な部分観測マルコフ決定過程（POMDP）下の環境をそれぞれ用いた。FAC は、いずれのシミュレーションにおいても、最適解や準最適解を見つけることができなかったが、学習の収束が見られた。また、個体間の適切な距離を保つことに正の報酬を与える場合を「群学習(Swarm Learning)」と、群行動を考慮しない場合を「単独学習（Individual Learning）」と呼び、シミュレーションの結果より、前者の学習性能が優れていることが明らかになった。

3.2 FN を用いた Q 学習型強化学習システム(FQ)

3.2.1 FQ の構成

ファジィネット(FN: Fuzzy net)を用いた Q 学習型強化学習システム(FQ)の構成及び環境との相互作用を図 3.18 に示す。ここで、「Fuzzy net」は 3.4 節で述べた自己組織化ファジィニューラルネットワーク(SOFNN)であり、「State-Action Value Function　$Q(\phi(\mathrm{x}(t)),\mathrm{A},\mathrm{w})$」は FN の出力（推論結果）に重み付けした「状態―行動価値関数」である。

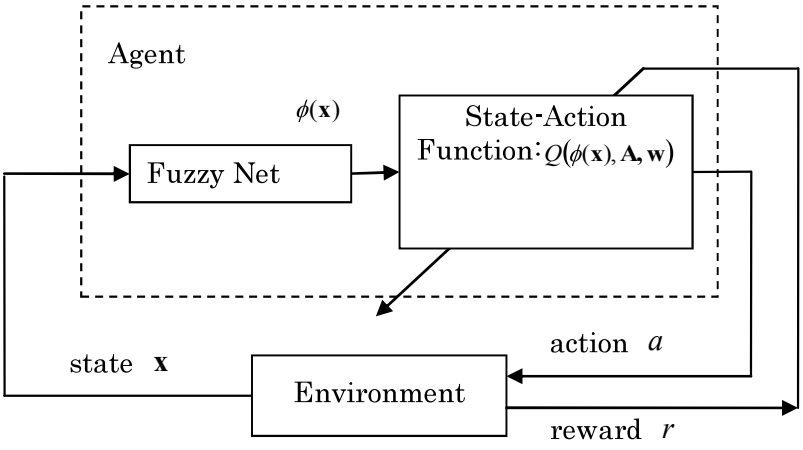


図 3.18 FQ を用いる知的個体と環境の相互作用

FQ の詳細な構成を図 3.19 に示す。まず環境から観測した状態が Fuzzy net 部に入力される。Fuzzy net 部では入力 $\mathbf{x}(t)$とメンバーシップ関数 $B_{im_i}^k(x_i(t)) = \exp\left\{-\frac{1}{2}\left(\frac{x_i(t)-c_{im_i}^k}{v_{im_i}^k}\right)^2\right\}$

（(2.8)式）からルールの適合度 $\phi_k(\mathrm{x}(t)) = \prod_{i=1}^{n} B_{im_i}^k(x_i(t))$ （(2.7)式）を計算し、そのファジィ推論の後件部である出力に結合荷重 $w_{kj}(k=1,2,...,K, j=1,2,...,J)$ (図 3.19 の結合線上の黒い点)を付け、非ファジィ化によって、状態―行動価値関数

$$Q(\phi(\mathrm{x}(t)), a_j, \mathrm{w}_j) = \frac{\sum_k w_{kj}\phi^k(\mathrm{x}(t))}{\sum_k \phi^k(\mathrm{x}(t))} \tag{3.9}$$

を構成する。ここで、$k=1,2,...,K$ はルール番号であり、$j=1,2,...,J$ は行動番号、$i=1,2,...,n$ は入力の要素番号である。エージェント（知的個体）は確率方策関数 $\pi(Q(\phi(\mathrm{x}(t)), a_j, \mathrm{w}_j))$の確率分布

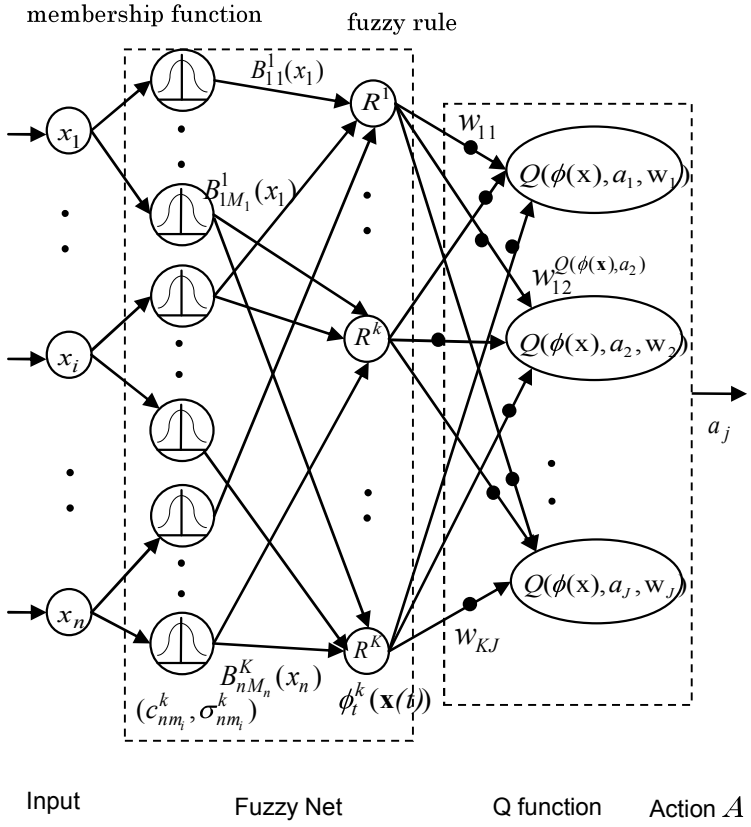


図 3.19　FQ の構成

$$p(a_t = a_j \mid x_t) = \frac{\exp(Q(\phi(\mathrm{x}(t)), a_t, \mathrm{w}_t)/T)}{\sum_{j=1}^{J} \exp(Q(\phi(\mathrm{x}(t)), a_j, \mathrm{w}_j)/T)} \tag{3.10}$$

に従って確率的に行動 a_j を選択する。ここで T>0 は温度定数というパラメータである。

3.2.2 FQ の学習則

任意の状態（入力パターン）に対して、正の報酬を最大限に獲得するよう、確率方策の修正は、(3.9)式にある結合加重 w_{kj} の修正によって間接的に行われる。また、w_{kj} の修正は、状態－行動価値関数 $Q(\phi(\mathrm{x}(t)), a_j, \mathrm{w}_j)$ の TD 誤差 $\varepsilon_{TD}^{Q}(\mathrm{x}(t))$ を用いて行う。

$$w_{kj}^{new} \leftarrow w_{kj}^{old} + \begin{cases} \alpha^{Q} \varepsilon_{TD}^{Q} \phi^k(\mathrm{x}(t)) & a_t = a_j \\ 0 & otherwise \end{cases} \tag{3.11}$$

但し、ここで $0 < \alpha^{Q} \le 1$ は学習率であり、TD 誤差 ε_{TD}^{Q} は以下のように定義される。

$$\varepsilon_{TD}^{Q} = r_t + \gamma \max_{a_{t+1} \in A} Q(\phi(\mathrm{x}(t+1)), a_{t+1}, \mathrm{w}_{t+1}) - Q(\phi(\mathrm{x}(t)), a_t, \mathrm{w}_t) \tag{3.12}$$

なお、r_t は状態 $\mathrm{x}(t)$ で行動 a_t によって状態 $\mathrm{x}(t+1)$ に到達する際に獲得した報酬であり、γ は割引率、$\max_{a_{t+1} \in A} Q(\phi(\mathrm{x}(t+1)), a_{t+1}, \mathrm{w}_{t+1})$ は次状態 $\mathrm{x}(t+1)$ の最大 Q 値である。

3.2.3 FQ の学習率

一般的に(3.11)式の学習率 α^{Q} を固定とする場合、学習回数の増加に伴って、更新量が振動してしまう現象が生じる。それを回避するため、学習回数の増加につれ、学習率を減衰させることが有効である。本論文では、Derhami ら[109]が提案した ALR 法（Adaptive Learning Rate）を学習率の調整に導入する。

Fuzzy Net のルール k の発火強度 $\phi^{k}(\mathbf{x}(t))$ の累積を Φ_t^k とし、状態 $\mathbf{x}(t)$ へのファジィ訪問価値を次式より計算する。

$$FV\left(\phi^{k}(\mathbf{x}(t))\right) = \frac{\sum_{k=1}^{K_t} \phi^{k}(\mathbf{x}(t)) \Phi_t^k}{\sum_{k=1}^{K_t} \Phi_t^k} \tag{3.13}$$

ここで $0 \le FV(\mathbf{x}(t)) \le 1$、$\Phi_t^k = \Phi_{t-1}^k + \phi^{k}(\mathbf{x}(t))$ で、$\Phi_{t=0}^k = 0$ とする。また、K_t は時刻 t における Fuzzy Net のルール数である。

(3.11)式の固定値の学習率 α^{Q} の代わりに、次式の適応的な学習率 $\alpha_t^{ALR}(\phi^{k}(\mathrm{x}(t)))$ を用いる。

$$\alpha_t^{ALR}(\phi^{k}(\mathrm{x}(t))) = \min\left(\frac{\alpha^{Q} K_t}{FV(\mathbf{x}(t))}, \alpha_{\max} \right) \tag{3.14}$$

ここで、α^{Q} は初期に設定された学習率（従来の固定値の学習率）、$\alpha_{\max}$ は α^{ALR} の上限値である。

ALR 法では、ある状態の訪問回数が多くなると、ファジィ訪問価値 $FV(\mathbf{x}(t))$ が増大し、学習率 $\alpha_t^{ALR}(\phi^{k}(\mathrm{x}(t)))$ が減少することとなり、学習過程に応じて適切な学習率を定めることができる。

3.2.4 FQ の群学習と単独学習

複数個体による目標探索問題において、個体間の距離を適切に保持するような行動を選

択されるように距離を考慮し、報酬を与える場合を FQ の「群学習（Swarm Learning)」と呼び、そうでない場合は FQ の「単独学習(Individual Learning)」と呼ぶ。個体間の距離の計算は(3.7)式、報酬の与え方は(3.8)式と同様である。

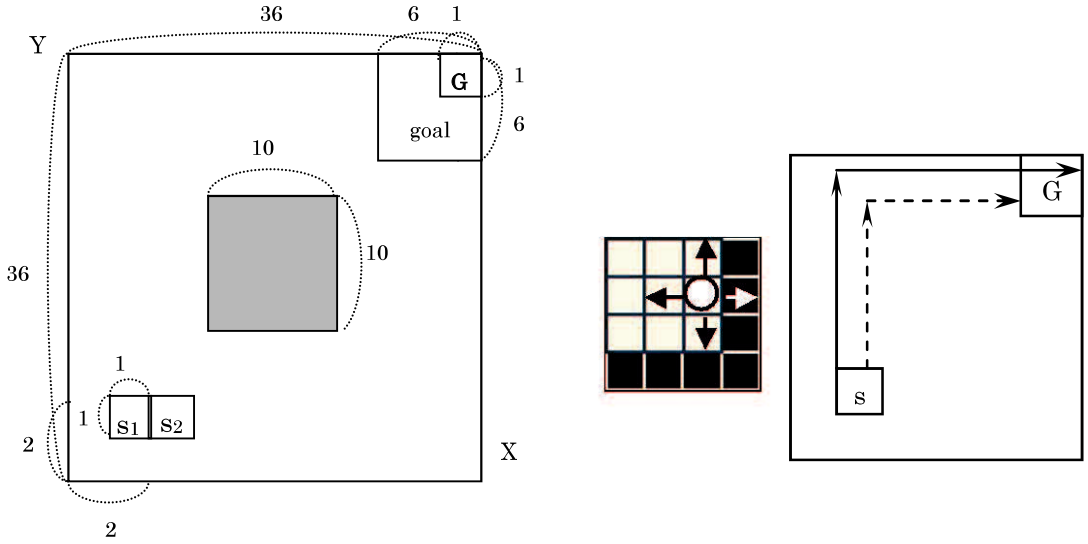
3.2.5 FQ の計算機シミュレーション

本章で述べた FQ を用いた POMDP 環境での目標探索シミュレーションを行い、FQ の有効性及び単独学習、群学習の優劣を明らかにする。

図 3.14(a)と同様な障害物が存在する探索環境を用いる（図 3.20(a)）。2 体の個体は(2,2)と(3,2)の出発点から、中央に 10x10 の障害物を回避し、(30,30)と(36,36)の間のゴールエリアを到達する最短経路を探索する問題設定である。各個体の行動は上下左右方向で、距離は 1 ステップ 1 マスで、計 4 個となる。また、観測する状態は上下左右 4 方向の近傍情報であり、4 次元ベクトルである（図 3.20(b)）。近傍情報とは、個体の 1 マス離れた位置の環境は通路やゴールエリアであれば 0、壁や障害物の場合は 1 の値のことを指す。すなわち、各個体が観測できる状態数は 2^4=16 個あることになる。個体が取り得る行動は上下左右方向 1 ステップ 1 マス計 4 個である。探索目標であるゴールエリアにおいて、「goal」と「G」で示される領域があり、それぞれの領域に到達したときの報酬が異なる。より深いところの G でより高い報酬値が与えられる。これは 2 体の個体が個体間に一定の距離を保ちながら群れを形成してゴールエリアへ到達させるためである。

図 3.20(c)は 1 体の個体がゴールエリアに到達する最短経路（破線）と準最短経路（実線）の例を示す。出発時の観測状態を(0, 0, 0, 0)とし、最短経路は、(0, 0, 0, 0)の状態が続く中で、個体が取る行動が「上」から「右」に変化し、同じ状態を観測したとしても異なる行動を取ることによって実現する。一方、個体が(0, 0, 0, 0)の状態遷移の途中、壁に沿う状態(1, 0, 0, 0)に変わってはじめて最適な行動が「上」から「右」に変化する場合、準最短路（実線）という準最適解しか得られない。観測状態が同じでも、実際の状態は異なることによって、異なる適切な行動を出力することができない問題は POMDP 環境の「エリアシング問題(aliasing problem)」と呼ばれる[7] [13] [15]。前章で述べた FAC が繰り返し探索によって個体の適応行動の学習が収束したことが認められたが、最適解や準最適解を見つけることができなかった。本章では、状態と行動（その組み合わせ）を共に評価する Q 学習方式によって、POMDP のエリアシング問題に対応することを試みる。

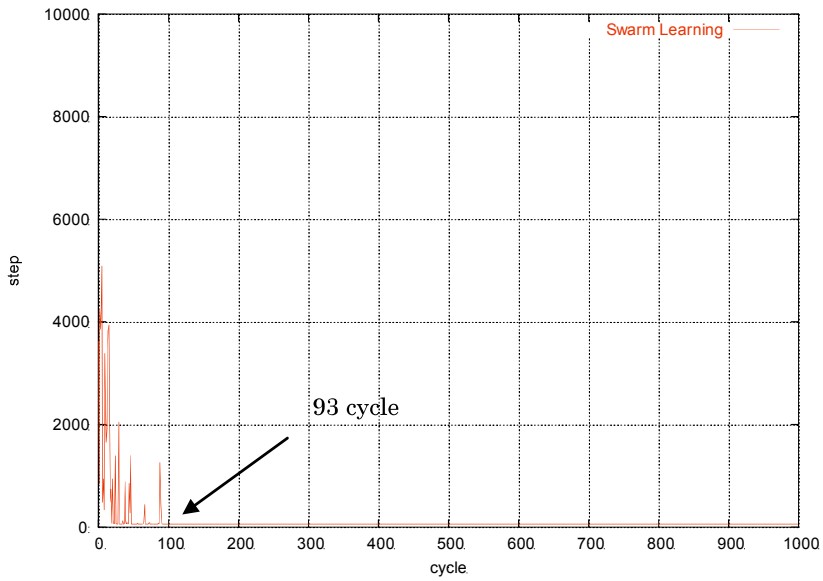
表 3.4 に FQ を用いた探索シミュレーションのパラメータを示す。図 3.21 は 2 体の個体が FQ を用いた学習による探索経路長の減少及び収束状況を示す。単独学習(Individual Learning)場合と群学習（Swarm Learning）の場合共に最適解の 56 step に達しなかったが、準最適解の 62 step に収束した。また、群学習の場合の収束が早く（93 試行（cycle））、収束後の安定性も単独学習の場合(333 試行)より優れていることが明らかになった。



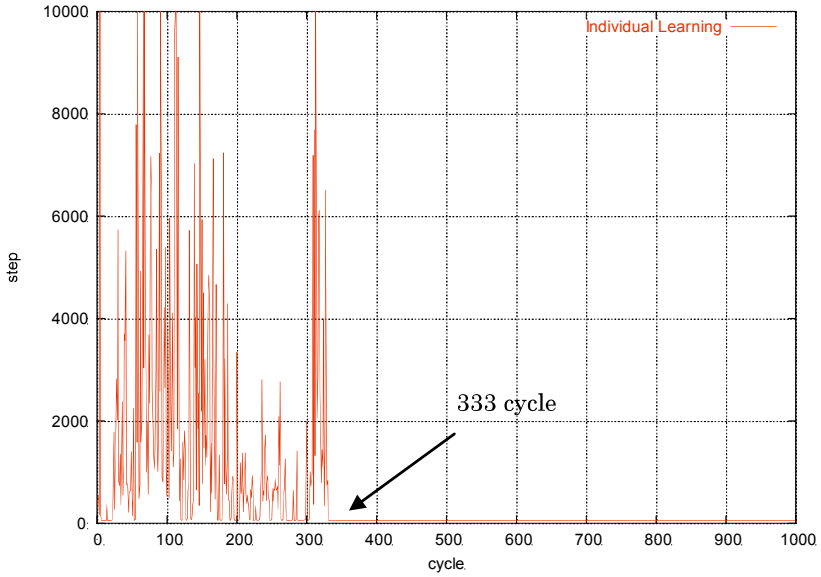
(a) 問題設定　　(b) 観測・行動方向　　(c) 最短経路と準最短経路

図 3.20　POMDP 下の未知環境探索問題

学習終了時の両個体の探索経路を図 3.22 に示す。群学習（図 3.22(a)）と単独学習（図 3.22(b)）の経路長はいずれも 62step のマンハッタン距離であったが、群学習の結果として、両個体が常に同じ経路を通ったことが分かる。



(a) 群学習の場合

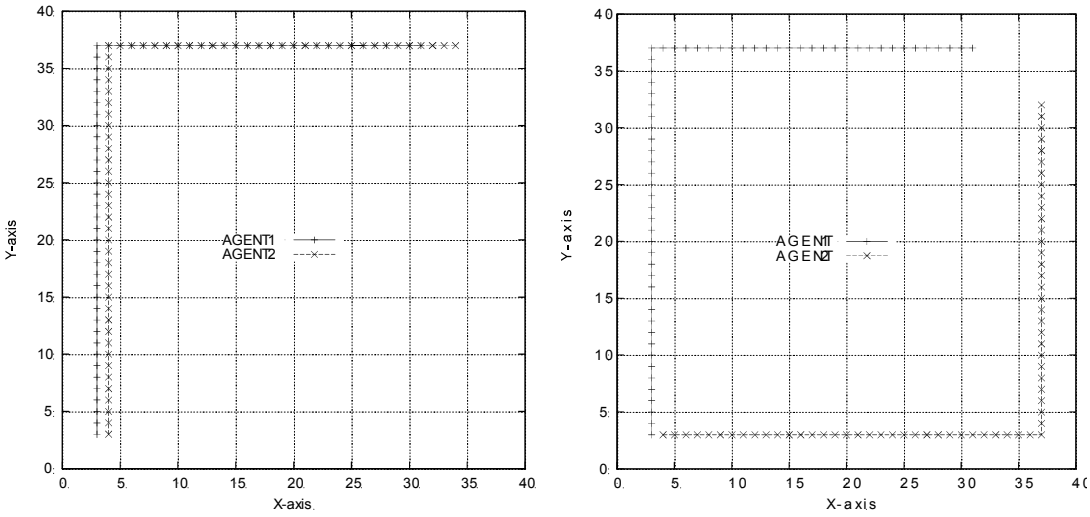


(b) 単独学習の場合

図 3.21 FQ の学習性能

FQ への 4 次元入力（上・下・左・右）に対し、Fuzzy net のファジィルールの増加の様子を図 3.23 に示す。いずれの次元とも入力値の 0 と 1 に対応する二つのメンバーシップ関数と、すべての状態を表す 16 個のファジィルールが生成されたことが分かる。

なお、1,000 回試行で探索した経路長の平均値は、群行動を考慮した群学習の場合は 117.6 step、群行動を考慮しない単独学習の場合は 661.2 step で、群学習の方が優れた学習性能を持つことが確認できた。



(a) 群学習の場合　　**(b) 単独学習の場合**

図 3.22 FQ の学習結果(探索経路)

表 3.4 POMDP 下の場合の FQ のパラメータ

記　述	符号	値
入力ベクトルの次元数	n	4
出力行動空間の次元数	J	4
メンバーシップ関数の広がり	σ^2	0.4
メンバーシップ関数の増殖閾値	F	0.4
Q への結合荷重の初期値	w_{kj}	1.0
学習率初期値	α^Q	0.001
学習率最大値	$\alpha_{\max}$	0.3
TD-error の減衰率	γ	0.8
確率方策における温度定数	T	0.8→0.2
ゴールエリア goal の報酬	r_{goal}	50.0
ゴールエリア G の報酬	r_{goal}^G	100.0
障害物に衝突する報酬	r_{crash}	-10.0
適切距離を保つ報酬	r_{swarm}	1.0
適切距離を保たない報酬	$r_{no\text{-}swarm}$	$-D$ (距離)
最小適切距離閾値	min_dis	1.5
最大適切距離閾値	max_dis	3.0

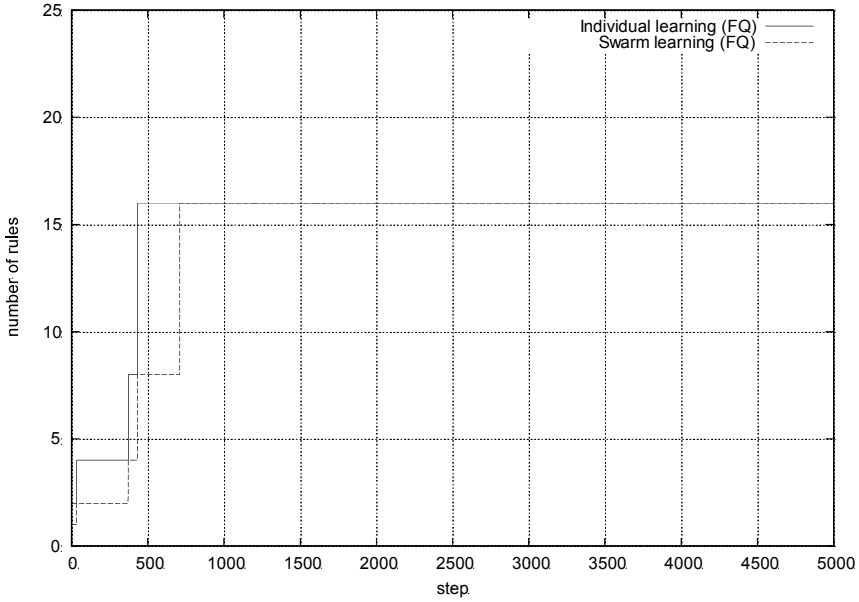
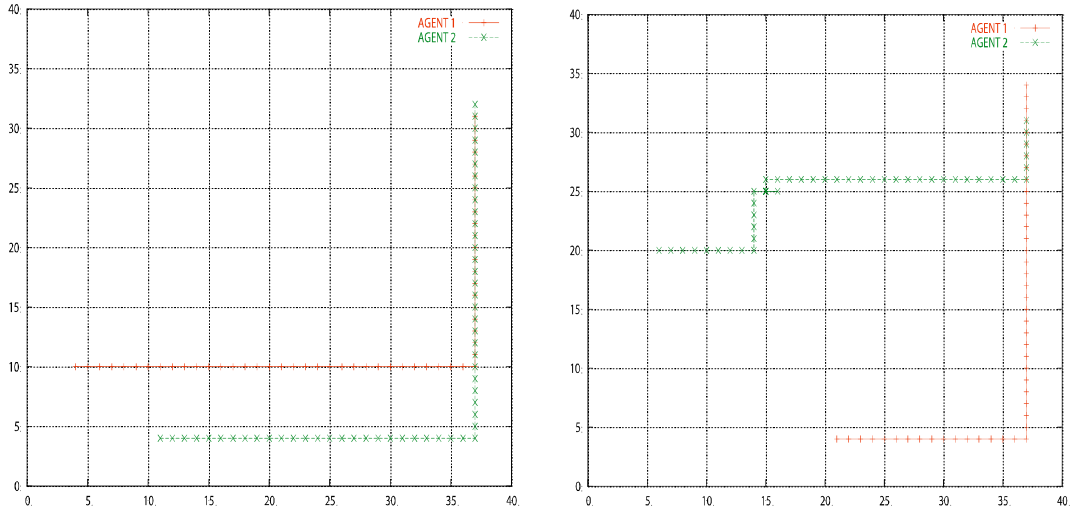


図 3.23 POMDP 環境を探索する FQ の Fuzzy net のルール数の変化



(a)　　(b)

図 3.24 POMDP 環境を探索する FQ の学習ロバスト性

図 3.24 は FQ の学習終了後、両個体の出発点を(2,2)と(2,3)から変更し、(3,10)と(10,4)にした場合(a)と、(5, 20)と(20,4)にした場合(ｂ)の探索軌跡を示す。いずれも個体間の距離を縮小しながら探索目標エリアに辿り着いたことが確認でき、FQ の学習ロバスト性が確認された。

3.2.6 本節のまとめ

本章では 3 章で述べた自己組織化ファジィニューラルネットワーク（SOFNN）を用いて、強化学習の Q 学習方式と融合し、新たなファジィニューラルネットワーク型強化学習システム FQ を提案した。また、FQ を用いた複数個体の目標探索問題のシミュレーション及びそれらの結果を示し、提案システムの有効性を確認することができた。目標探索問題の環境設定について、障害物が存在し、個体の近傍環境のみが観測可能な部分観測マルコフ決定過程（POMDP）の環境を用いた。前章で提案された FAC に比べ、FQ の学習性能が準最適解を見つけることができるまで大幅に高まった。また、個体間の適切な距離を保つことに正の報酬を与える「群学習(Swarm Learning)」は、群行動を考慮しない「単独学習（Individual Learning）」より、学習性能が優れていることが明らかになった。

3.3 FN を用いた Sarsa 学習型強化学習システム(FS)

3.3.1 FS の構成

「Sarsa」は状態 s 、行動 a、次状態 s 、次行動 a の綴りを用いた強化学習法で、連続した二つの状態と行動の価値（報酬）を利用しているため、Q 学習より「用心深い」経路を選ぶことができる [15]。ここで、前章の FQ の Q 学習の代わりに、Sarsa 学習をニューロファジィ型強化学習システムに導入し、新たな強化学習システム FS を構築する。

図 3.25 は FS を用いる知的個体と環境の相互作用を示している。個体(Agent)が、時刻 *t*

及び t+1 における状態 $\phi(\mathrm{x}(t))$ 及び $\phi(\mathrm{x}(t+1))$ を観測し、それぞれの状態―行動価値関数 $Q(\phi(\mathrm{x}(t)),a_t,\mathrm{w}_t)$ と $Q(\phi(\mathrm{x}(t+1)),a_{t+1},\mathrm{w}_{t+1})$ をネットワークの出力によって計算し、それらの Q 値を用いた確率政策によって行動を実行する。

図 3.26 は FS の詳細構成を示している。TD 学習による Q 値の修正を考慮し、FS の入力が時刻 t と時刻 t+1 の場合を同時に表現した図 3.26 は、図 3.19 に示した FQ より、やや複雑な形となっている。

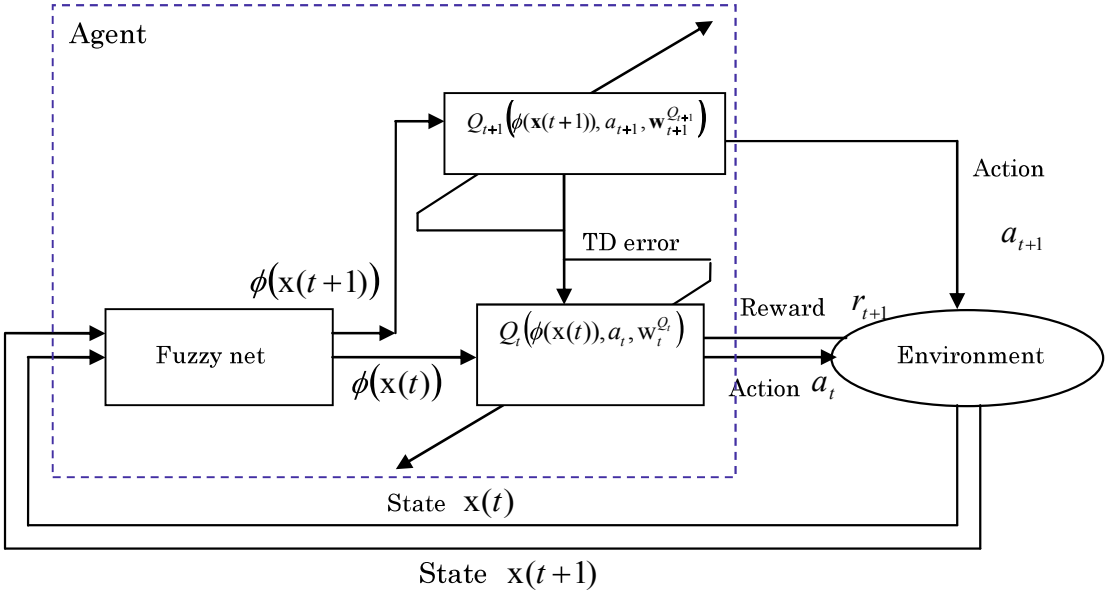


図 3.25 FS を用いる知的個体と環境の相互作用

3.3.2 FS の学習則

FS では、従来の Sarsa 学習アルゴリズム(文献[15]及び付録参照)と同様に「TD 誤差を用いて Q 値を修正する」学習則を用いるが、その修正は、結合加重の修正によって間接的に修正する。すなわち、各ルールと各状態―行動価値関数 Q_t, Q_{t+1} の結合加重 $w_{kj}^{Q_t}, w_{kj}^{Q_{t+1}}$ は、以下のように修正する：

$$w_{kj}^{Q_t} \leftarrow w_{kj}^{Q_t} + \begin{cases} \alpha^{sarsa}\varepsilon_{TD}^{sarsa}\phi_k(\mathrm{x}(t))/\sum_{k=1}^{K}\phi_k(\mathrm{x}(t)) & a_t = a_j \\ 0 & otherwise \end{cases} \tag{3.15}$$

$$w_{kj}^{Q_{t+1}} \leftarrow w_{kj}^{Q_{t+1}} + \begin{cases} \alpha^{sarsa}\varepsilon_{TD}^{sarsa}\phi_k(\mathrm{x}(t+1))/\sum_{k=1}^{K}\phi_k(\mathrm{x}(t+1)) & a_{t+1} = a_j \\ 0 & otherwise \end{cases} \tag{3.16}$$

ここで、$0 < \alpha^{sarsa} \le 1$ は学習率であり、TD 誤差 ε_{TD}^{sarsa} は以下のように定義される。

$$\varepsilon_{TD}^{Q_{t+1}} = r_t + \gamma\, Q(\phi(\mathrm{x}(t+1)), a_{t+1}, \mathrm{w}_{t+1}^{Q_{t+1}}) - Q(\phi(\mathrm{x}(t)), a_t, \mathrm{w}_t^{Q_t}) \tag{3.17}$$

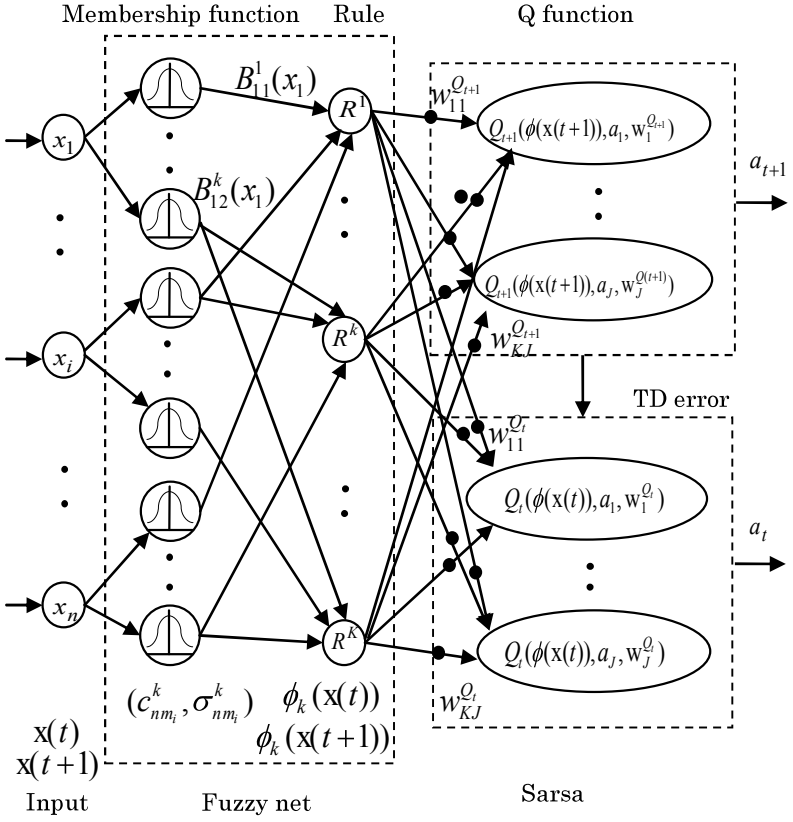


図 3.26　FS の詳細構成

なお、r_t は状態 $\mathrm{x}(t)$ で行動 a_t によって状態 $\mathrm{x}(t+1)$ に到達する際に獲得した報酬であり、γ は割引率である。

Q 学習型強化学習システムの学習則である(3.12)式に比べ、Sarsa 学習型の学習則(3.17)式は次状態の最大値を用いず、次状態で決定された行動の値 Q_t, Q_{t+1} を用いて TD 誤差を計算する。よって、Sarsa 学習は 2 ステップの状態―行動価値関数を用いるため、Q 学習に比べ、「より用心深い」行動選択方策を求める[15]。

なお、FS では、状態―行動価値関数の構成は FQ の(3.9)式と同様であるが、(3.18)式と(3.19)式のように、現状態と次状態の入力情報を共に用いる。

$$Q(\phi(\mathrm{x}(t)), a_j, \mathrm{w}_j^{Q_t}) = \frac{\sum_k w_{kj}^{Q_t} \phi^k(\mathrm{x}(t))}{\sum_k \phi^k(\mathrm{x}(t))} \tag{3.18}$$

$$Q(\phi(\mathrm{x}(t+1)), a_j, \mathrm{w}_j^{Q_{t+1}}) = \frac{\sum_k w_{kj}^{Q_{t+1}} \phi^k(\mathrm{x}(t+1))}{\sum_k \phi^k(\mathrm{x}(t+1))} \tag{3.19}$$

ここで、$k=1,2,...,K$ はルール番号であり、$j=1,2,...,J$ は行動番号、$i=1,2,...,n$ は入力の要素番号である。現状態 x(t)と次状態 x(t+1)における確率方策関数 π(Q_t)と π(Q_{t+1})は、それぞれ、以下の確率に従って行動 a_t, a_{t+1} を選択する。

$$p(a_t = a_j \mid \mathrm{x}(t)) = \frac{\exp(Q(\phi(\mathrm{x}(t)), a_t, \mathrm{w}_j^{Q_t})/T)}{\sum_{j=1}^{J} \exp(Q(\phi(\mathrm{x}(t)), a_j, \mathrm{w}_j^{Q_t})/T)} \tag{3.20}$$

$$p(a_{t+1} = a_j \mid \mathrm{x}(t+1)) = \frac{\exp(Q(\phi(\mathrm{x}(t+1)), a_{t+1}, \mathrm{w}_j^{Q_{t+1}})/T)}{\sum_{j=1}^{J} \exp(Q(\phi(\mathrm{x}(t+1)), a_j, \mathrm{w}_j^{Q_{t+1}})/T)} \tag{3.21}$$

ここで、T>0 は温度定数というパラメータである。

学習の収束性を高めるため、(3.15)式と(3.16)式にある固定の学習率 α^{Sarsa} の代わりに前節の FQ と同様に、適応的な学習率 $\alpha_t^{ALR}(\phi^k(\mathrm{x}(t)))$ を用いる。

$$\alpha_t^{ALR}(\phi^k(\mathrm{x}(t))) = \min\left(\frac{\alpha^{Sarsa} K_t}{FV(\mathbf{x}(t))}, \alpha_{\max}\right) \tag{3.22}$$

ここで、α^{Sarsa} は初期に設定された学習率（従来の固定値の学習率）、$\alpha_{\max}$ は α^{ALR} の上限値である。メタパラメータ $0 \le FV(\mathbf{x}(t)) \le 1$ は、(3.13)式で定義されている。

3.3.3 FS の計算機シミュレーション

本章で述べた FS を用いた POMDP 環境での目標探索シミュレーションを行い、FS の有効性及び単独学習、群学習の優劣を明らかにする。

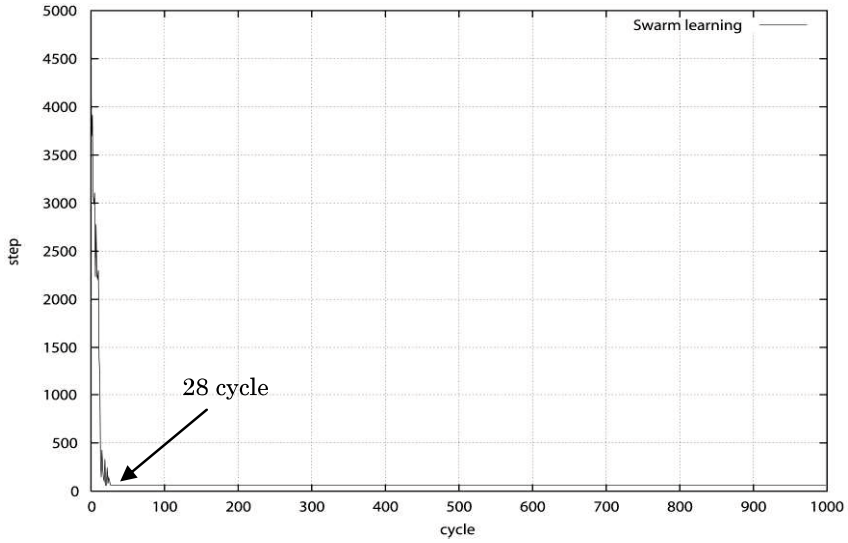
図 3.20(a)と同様な障害物が存在する探索環境を用いる。また、すべての問題設定は前節のシミュレーションと同様である。

表 3.5 に FS を用いた探索シミュレーションのパラメータを示す。図 3.27 は 2 体の個体が FS を用いた学習による探索経路長の減少及び収束状況を示す。単独学習(Individual Learning)場合と群学習（Swarm Learning）の場合共に最適解の 56step に達しなかったが、準最適解の 62step に収束した。また、図 3.27(a)と図 3.27(b)によって、群学習の収束速度（28 回目試行(cycle)から）は単独学習（407 回目試行(cycle)から）より、大幅に早まっていることが分かる。群学習の収束安定性は FQ の場合（図 3.21）に比べ、優れていることが明らかになった。

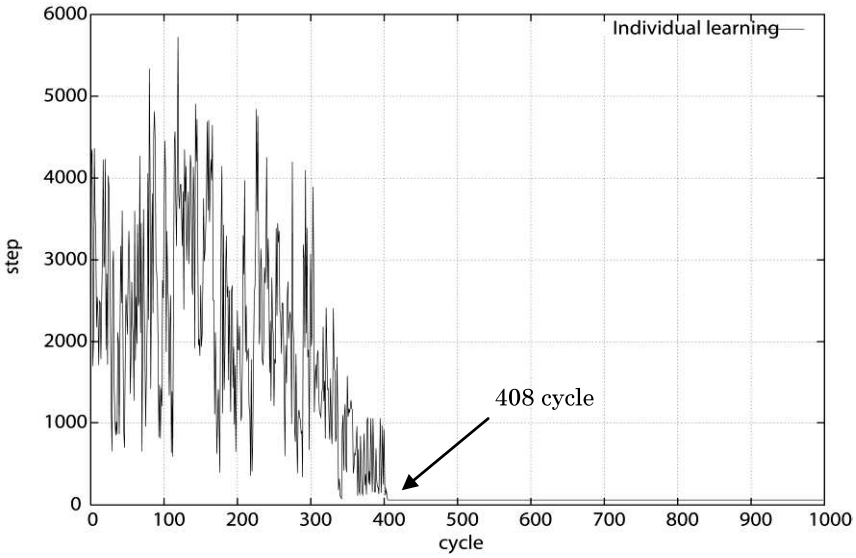
学習終了時の両個体の探索経路を図 3.28 に示す。群学習（図 3.28(a)）と単独学習（図 3.28(b)）の経路長はいずれも 62step のマンハッタン距離であったが、群学習の結果として、両個体が同じ経路を通っていることが分かる。

図 3.29 は FS への 4 次元入力（上下左右）に対し、Fuzzy net のファジィルールの増加の様子を示している。いずれの次元とも入力値の 0 と 1 に対応する二つのメンバーシップ関数と、すべての状態を表す 16 個のファジィルールが生成されたことが分かる。また、Fuzzy

Net の自己増殖の速さについて、群学習（破線）の方が単独学習（実線）より早く、それぞれ 2,438、8,544 ステップ目からファジィルールが完全に構築されていることが分かる。この現象も群探索の場合の優位性を裏付ける。

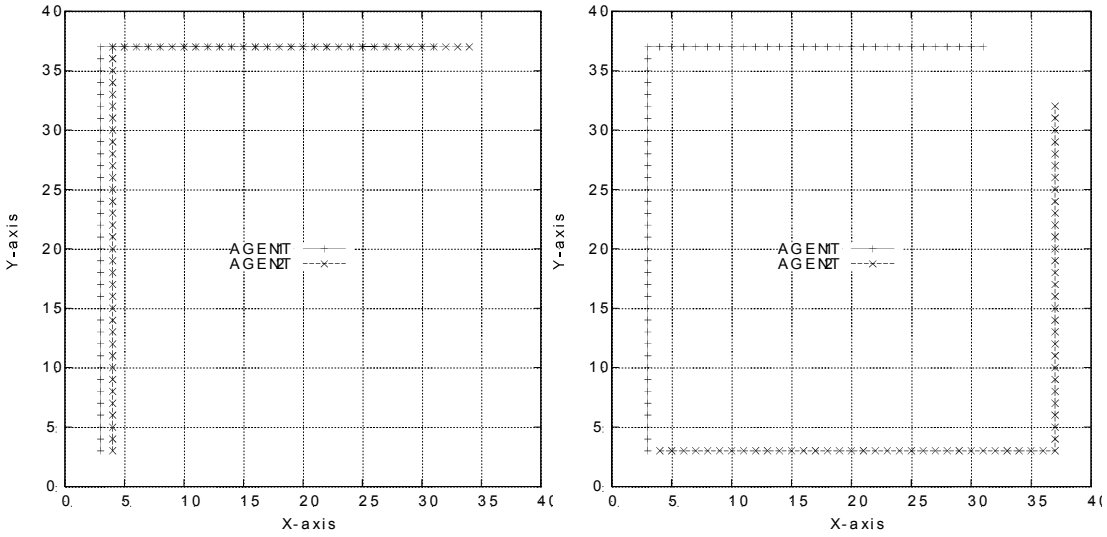


(a) 群学習の場合



(b) 単独学習の場合

図 3.27 FS の学習性能



(a) 群学習の場合　　**(b) 単独学習の場合**

図 3.28 FS の学習結果(探索経路)

表 3.5 POMDP の場合の FS のパラメータ

記　述	符　号	値
入力ベクトルの次元数	n	4
出力行動空間の次元数	J	4
メンバーシップ関数の広がり	σ^2	0.4
メンバーシップ関数の増殖閾値	F	0.4
Q_t, Q_{t+1} への結合荷重の初期値	$w_{kj}^{Q_t}, w_{kj}^{Q_{t+1}}$	1.0
学習率初期値	α^{Sarsa}	0.001
学習率最大値	$\alpha_{\max}$	0.3
TD-error の減衰率	γ	0.8
確率方策における温度定数	T	0.8→0.2
ゴールエリア goal の報酬	r_{goal}	50.0
ゴールエリア G の報酬	r_{goal}^{G}	100.0
障害物に衝突する報酬	r_{crash}	-1.0
適切距離を保つ報酬	r_{swarm}	1.0
適切距離を保たない報酬	$r_{no\text{-}swarm}$	-D (距離)
最小適切距離閾値	min_dis	1.5
最大適切距離閾値	max_dis	5.0

なお、1,000 回試行で探索した経路長の平均値は、群行動を考慮した群学習の場合は 97.6 step/cycle、群行動を考慮しない単独学習の場合は 885.4 step/cycle で、群学習の方が優れた学習性能を持つことが確認できた。

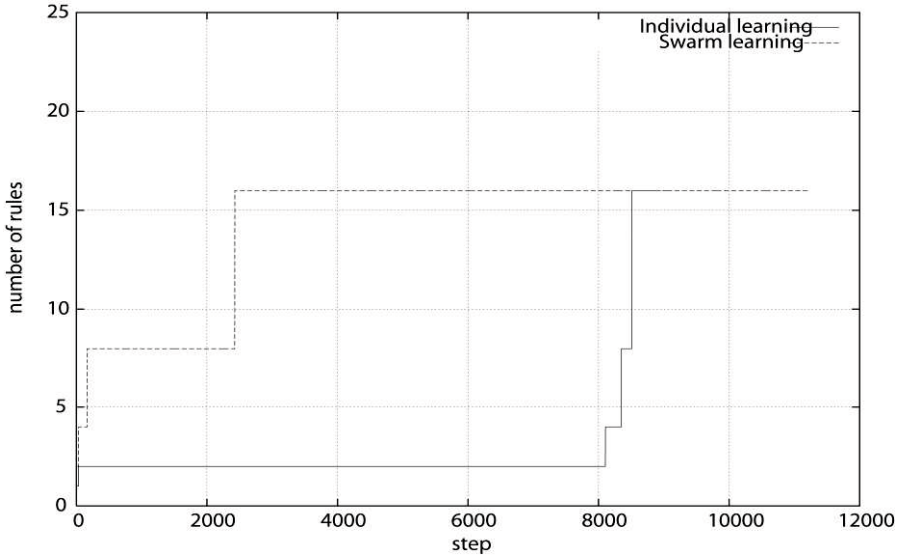
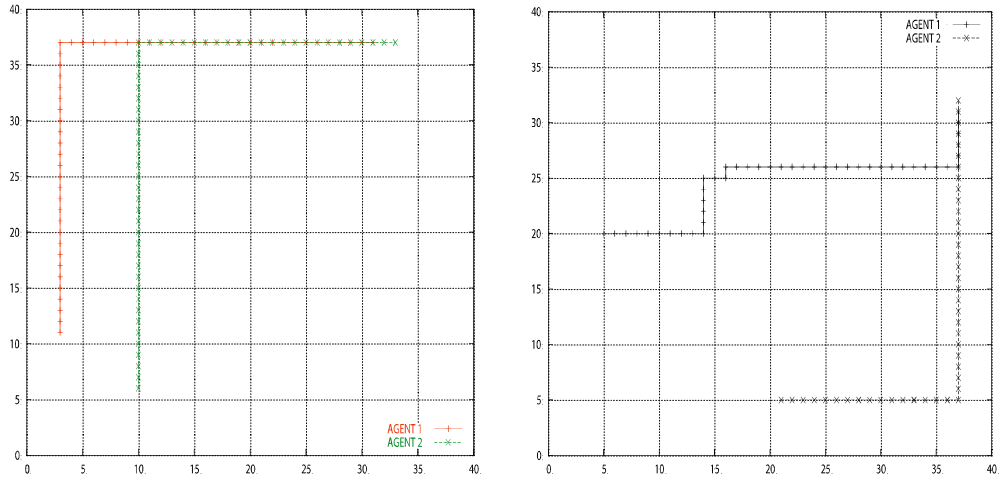


図 3.29 POMDP 環境を探索する FS の Fuzzy net のルール数の変化



(a)　　**(b)**

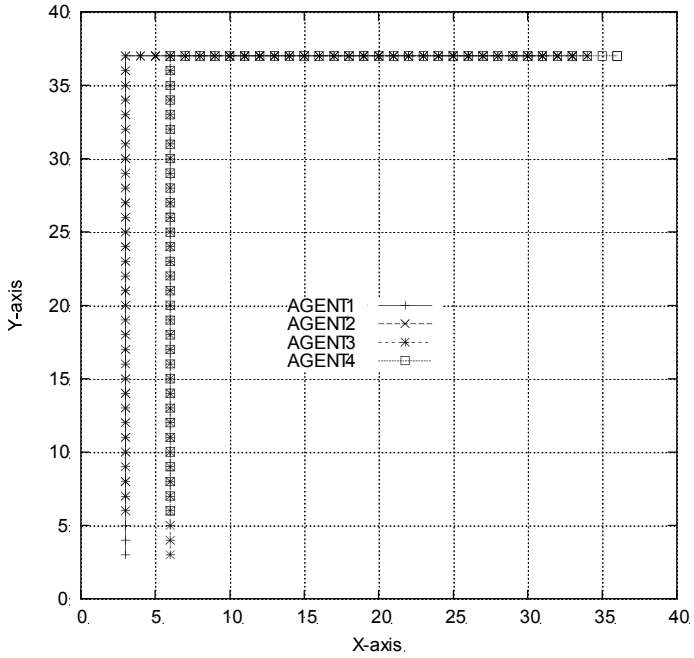
図 3.30 POMDP 環境を探索する FS の学習ロバスト性

図 3.30 は FS の学習終了後、両個体の出発点を(2,2)と(2,3)から変更し、(3,10)と(10,5)にした場合(a)と、(4, 20)と(20,5)にした場合(b)の探索軌跡を示す。いずれも個体間の距離を縮小しながら探索目標エリアに辿り着いていることが確認でき、FS の学習ロバスト性が確認された。

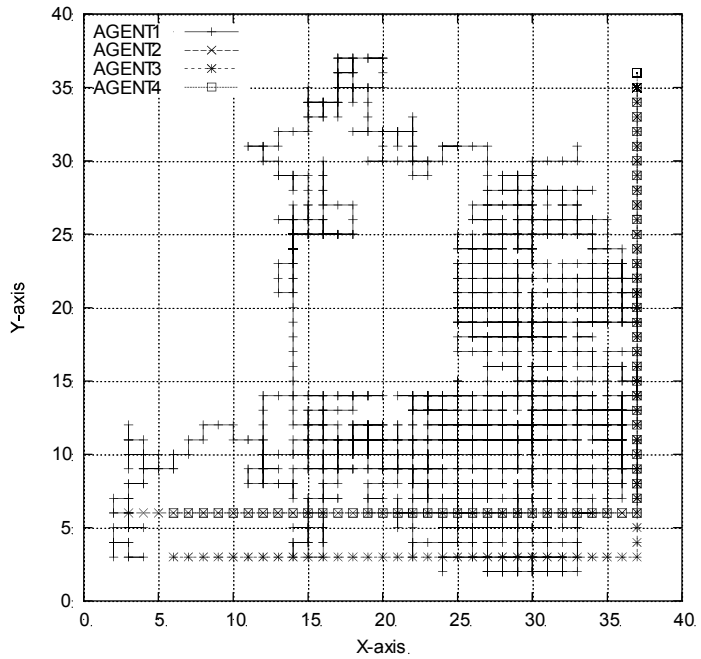
個体数が増加した場合のシミュレーションも行い、提案システムの性能を検証した。図 3.31 は知的個体 4 体で，図 3.20(a)の環境で出発点(3, 3), (6, 3), (3, 6), (6, 6)の場合の学習後の経路を示す。収束性能が良くない単独探索（図 3.31 (b)）より，群れを形成しながら目標地

点へ向かった場合（図 3.31(a)）の方が経路長が短く、効率が高いということが分かる。

図 3.32 は個体数が異なる場合、単独学習と群学習の学習コスト（経路長が収束するまでの 10 試行平均 step 数）を比較している。個体数が増えるほど、群学習の有効性がより顕著になることが明らかになった。



(a)　群学習の場合



(b)　単独学習の場合

図 3.31　個体 4 体の FS の学習結果(探索経路)

図 3.32 障害物ありの環境で異なる個体数の場合の FS の学習コストの比較

3.3.4 本節のまとめ

本節では前章で述べた自己組織化ファジィニューラルネットワーク（SOFNN）を用いて、強化学習の Sarsa 学習方式と融合し、新たなファジィニューラルネットワーク型強化学習システム FS を提案した。また、FS を用いた複数個体の目標探索問題のシミュレーション及びそれらの結果を示し、提案システムの有効性を確認することができた。目標探索問題の環境設定について、障害物が存在し、個体の近傍環境のみが観測可能な部分観測マルコフ決定過程（POMDP）の環境を用いた。前節で提案された Q 学習型強化学習システム FQ に比べ、FS の学習の収束安定性が高まった。また、個体間の適切な距離を保つことに正の報酬を与える「群学習(Swarm Learning)」は、群行動を考慮しない「単独学習（Individual Learning）」より、学習性能が優れていることが明らかになった。

第 4 章

考察

第 2 章で述べた自己組織化ファジィニューラルネットワーク（SOFNN）と従来の強化学習方式を融合し Actor-Critic 型強化学習システム(FAC)、Q 学習型強化学習システム（FQ）、及び Sarsa 学習型強化学習システム（FS）をそれぞれ第 3 章で構築し、また、目標探索問題のシミュレーション及びそれらの結果によって、各提案強化学習システムの有効性及び問題点を明らかにした。

本章では、まず、4.1 節では、提案法の FAC、FQ、FS と、従来の強化学習手法 Q 学習、Sarsa 学習、SGA 学習のシミュレーション結果を比較し、各手法の特徴について述べる。

次に、4.2 節では、確率方策関数におけるパラメータ（温度定数）による学習性能への影響を考察する。

更に、4.3 節では、学習則におけるパラメータ（学習率）の最適化による学習性能の向上をシミュレーションの結果を用いて明らかにする。

4.1 提案法と従来法のシミュレーション結果の比較

障害物のある未知環境における目標探索問題のシミュレーションは図 3.20 に示す環境設定で行い、前章で提案した FAC、FQ、FS を用いて、従来の強化学習手法である Q 学習 [15][28] と、Sarsa 学習 [15][28][31]と SGA 学習[46]の学習性能を比較する。

シミュレーションにおいては、個体が観測する状態は近傍 4 方向（上下左右）の値 0(通路)と 1（壁）によって構成され、総状態数は計 2^4=16 個存在し、「同じ観測状態でも実際の位置によって異なる適切な行動を取るべき」という不完全知覚問題が存在し、個体の状態遷移過程は、不完全観測マルコフ決定過程（POMDP）に属する。また、個体の行動は上下左右 4 方向の 1 ステップ 1 マスの 4 つとする。

4.1.1 ランダム探索の場合

個体が 2 体で、学習を用いず、ランダムに目標を探索し、2 体ともに目標領域に進入するときに探索が成功するシミュレーション結果を図 4.1 と図 4.2 に示す。

図 4.1 は 10 回シミュレーションを実行し、各シミュレーションにおいて、1,000 試行

(cycle)を行った場合の平均探索ステップ数(step)を示す。各試行のステップ数は試行回数の増加による変化はランダムであり、全体の平均ステップ数は 5,788.9 であった。

図 4.2 は探索終了時の個体の軌跡を表し、ランダムな探索の様子が明らかとなっている。

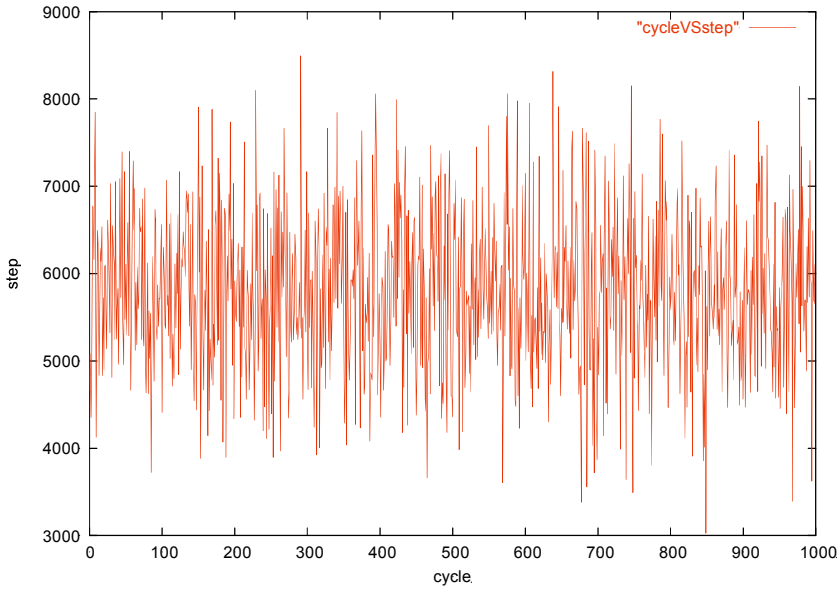


図 4.1 ランダム探索(学習なし)場合の経路長の変化

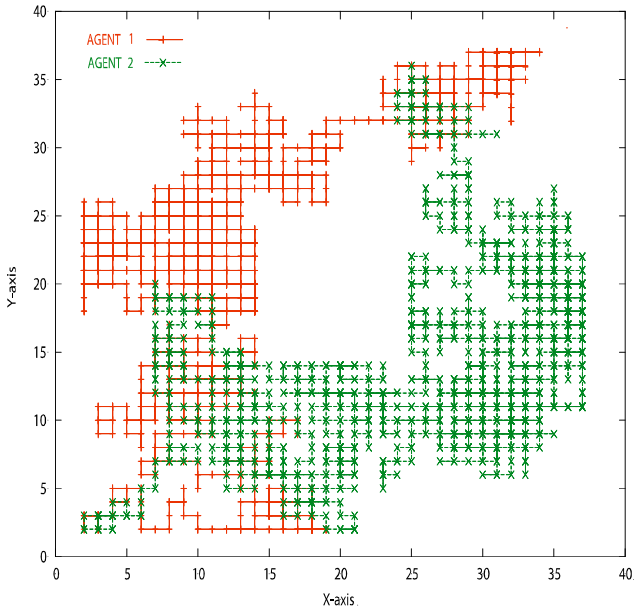


図 4.2 ランダム探索（2 体単独）の終了時の探索軌跡

4.1.2 従来法 1: Q 学習の場合

ランダム探索シミュレーションと同様な設定で、従来の Q 学習[15][28][31](付録 A 参照)を用いた個体 2 体の探索シミュレーション結果を図 4.3 と図 4.4 に示す。

図 4.3 は 10 回シミュレーションを実行し、各シミュレーションにおいて、1,000 試行(cycle)を行った場合の平均探索ステップ数(step)を示す。各試行のステップ数は、試行回数の増加につれ、減少し、収束する変化が見られる。収束時(1,000cycle)の平均経路長は群学習の場合 3,048.0 step で、単独学習の場合は 6,024.0step であった。

なお、1,000 試行の平均探索ステップ数は、群学習と単独学習それぞれ 3,147.4 と 6,041.1 step/cycle で、群学習の学習性能が比較的優れていることが確認できた。

図 4.4 は従来の Q 学習を用いて両個体の学習終了時に得られた経路を示す。図 4.4(a)に示した群学習の準最適解である 62 ステップのルートを発見することができたが、単独学習の場合は図 4.4(b)に示されたように、2 体個体は同一経路を経て目標エリアに到達することができなかった。

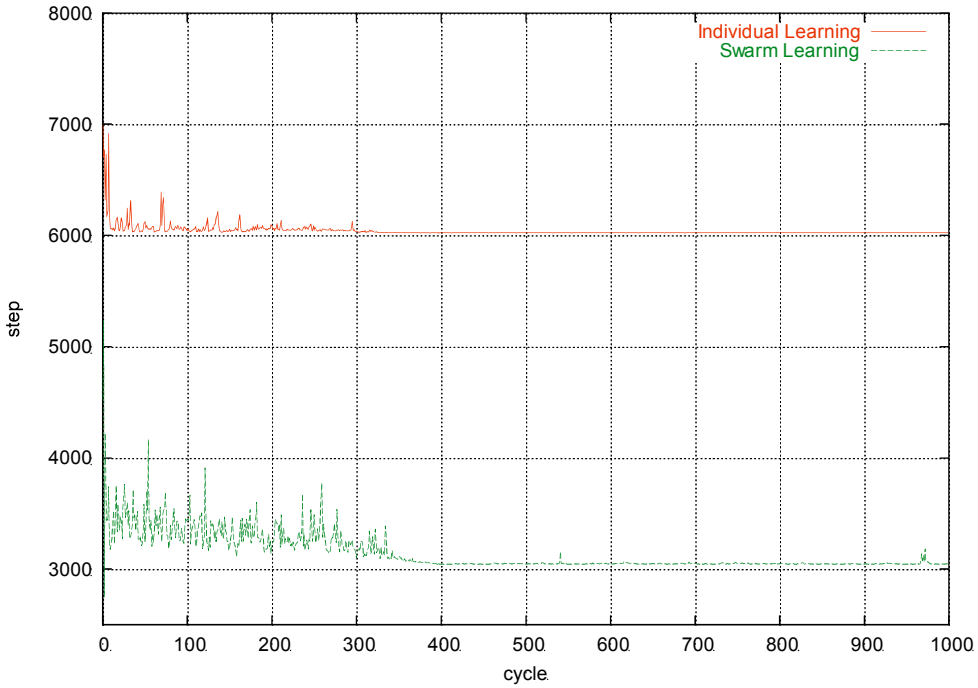
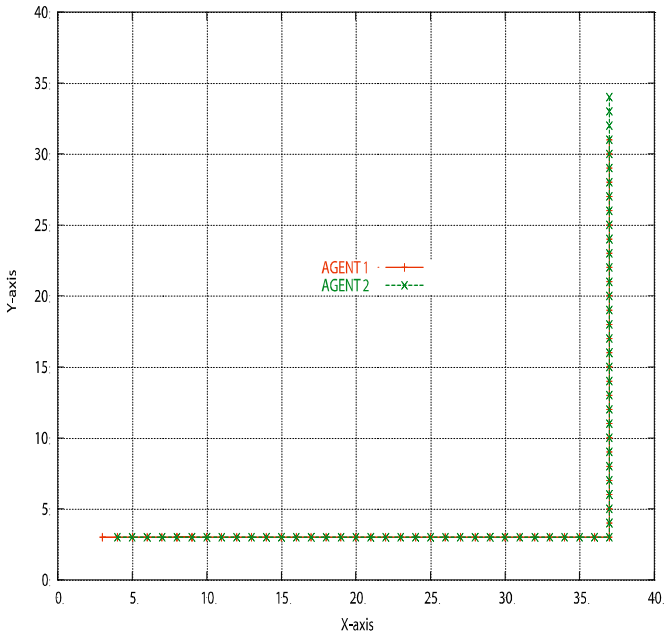
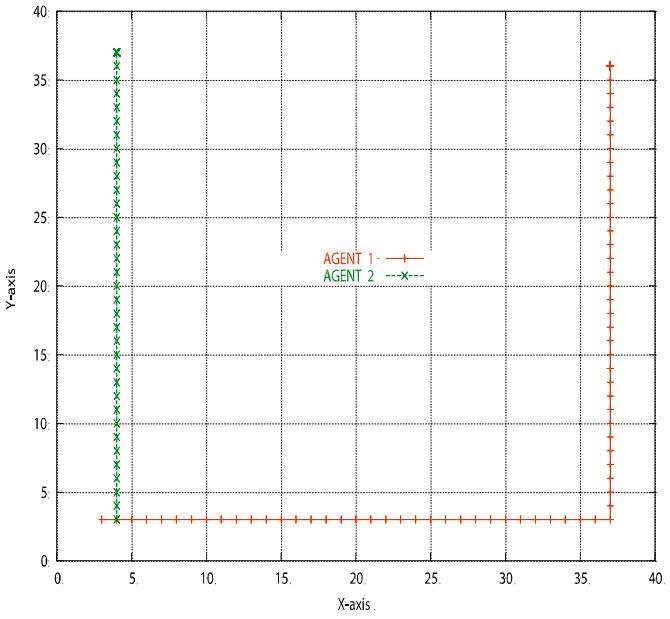


図 4.3 従来法 1: Q 学習を用いた探索における経路長の変化



(a)　群学習の場合



(b)　単独学習の場合

図 4.4 従来法 1:　Q 学習を用いた学習終了時の探索軌跡

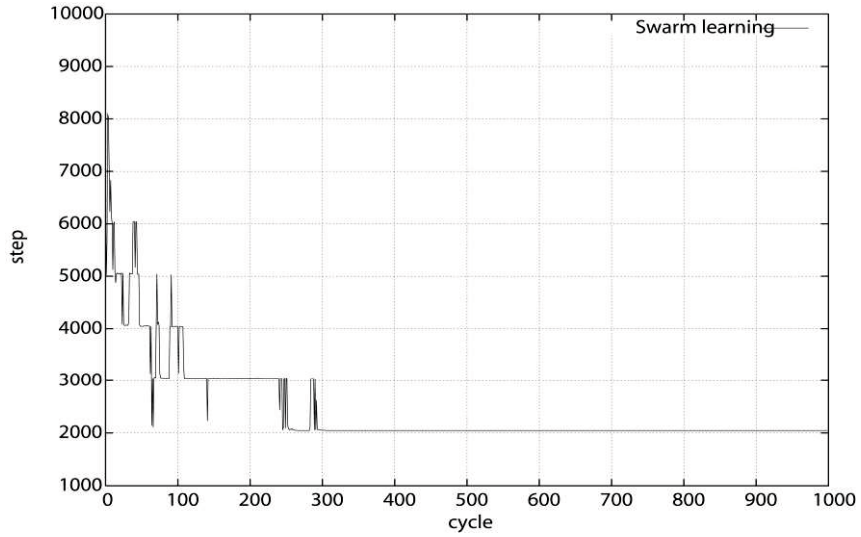
4.1.3 従来法 2: Sarsa 学習の場合

従来の Sarsa 学習[15][29][31](付録 B 参照)を用いた個体 2 体の探索シミュレーションの結果を図 4.5 と図 4.6 に示す。

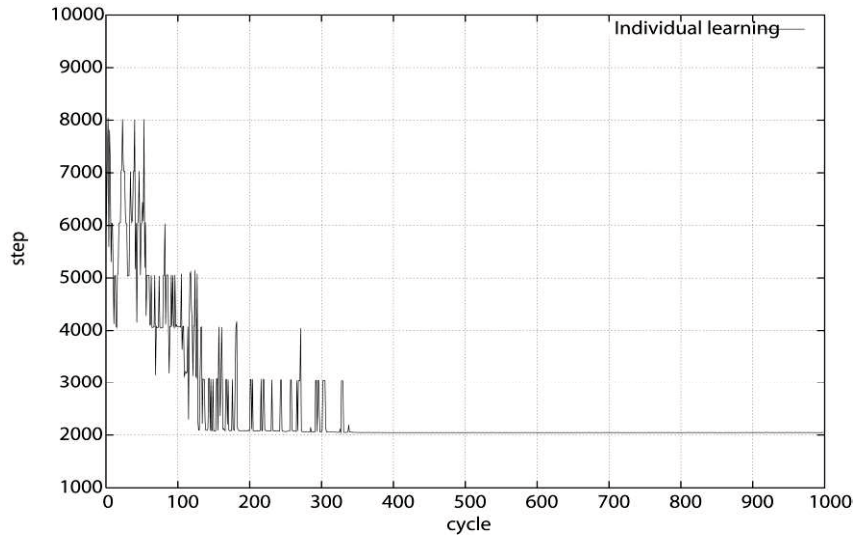
図 4.5 は 10 回シミュレーションを実行し、各シミュレーションにおいて、1,000 試行

(cycle)を行った場合の平均探索ステップ数(step)を示す。各試行のステップ数は、試行回数の増加につれ、減少し、収束する変化が見られる。収束時(1,000 cycle)の経路長は群学習(図 4.5 (a))の場合と、単独学習の場合(図 4.5 (b))ともに 2,048.0 step であった。なお、1,000 試行の平均ステップ数はそれぞれ 2,450.1 と 2,494.6 step/cycle で、群学習の学習性能が僅かであるが、優れていることが示された。

図 4.6 は従来の Sarsa 学習を用いて両個体の学習終了時に得られた経路を示す。2 体個体が共に目標エリアに到達することができたが、単独学習の場合は 2 体個体の経路は異なっていた。

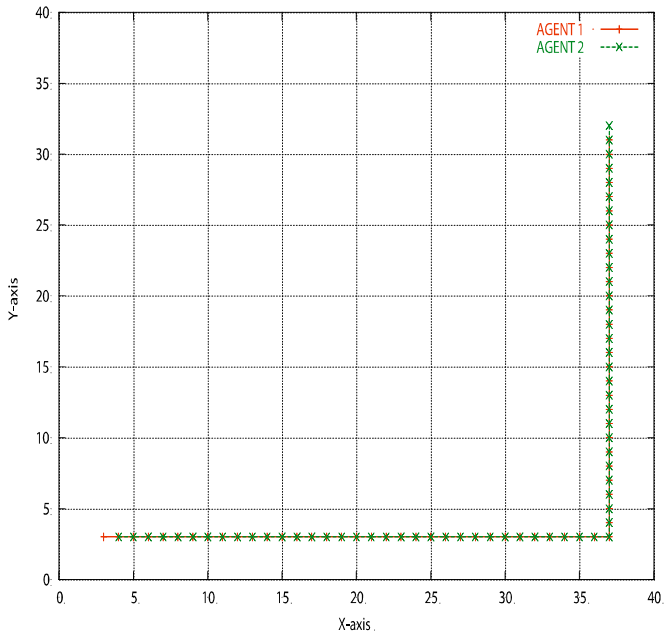


(a)　群学習の場合

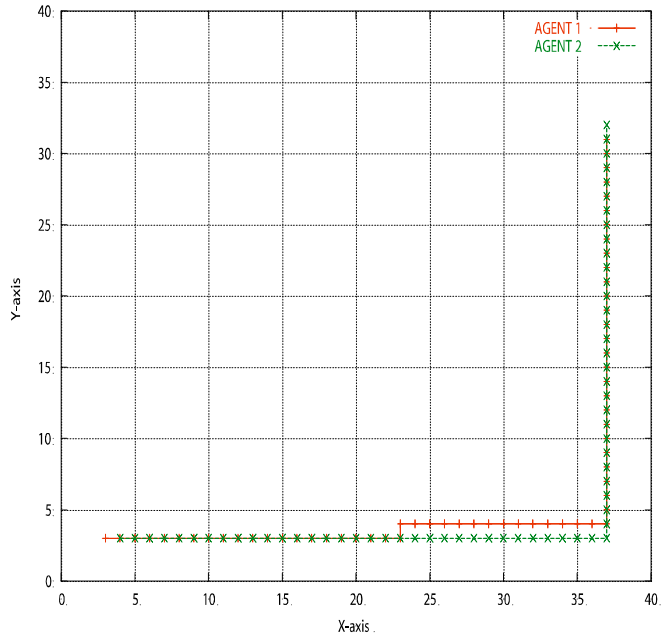


(b)　単独学習の場合

図 4.5 従来法 2: Sarsa 学習を用いた探索における経路長の変化



(a) 群学習の場合



(b) 単独学習の場合

図 4.6 従来法 2: Sarsa 学習を用いた学習終了時の探索軌跡

4.1.4 従来法 3: SGA 学習の場合

自己組織化ファジィニューラルネットワーク(SOFNN)[44]を用いた強化学習システムで梅迫らの「自己組織化型ファジィ強化学習システム」が提案されている[46]。ファジィネッ

トの出力に重み付け、行動選択方策の確率関数のパラメータと結合し、Stochastic Gradient Ascent (SGA)学習[56]によって行動方策の学習が実現されている。

従来の SGA 学習を用いた自己組織化型ファジィ強化学習システム[46](付録 C 参照)による個体数 2 の探索シミュレーションを行い、その結果を図 4.7 と図 4.8 に示す。

図 4.7 は 10 回のシミュレーションを行い、各シミュレーションにおいて、1,000 試行(cycle)を行った場合の平均経路長（1 試行当たりの探索ステップ数）を示す。各試行のステップ数は、試行回数の増加につれ、減少の後、収束が見られるが、やや振動的で、安定性に欠ける。1,000 試行の経路長は群学習の場合と、単独学習の場合それぞれ 232step と、327step であった。なお、1,000 試行の平均ステップ数はそれぞれ 733.2 と 1,045.5 step/cycle で、群学習の学習性能が優れていることが示された。

図 4.8 は従来の SGA 学習を用いて両個体の学習終了時に得られた経路を示す。2 体の個体が共に目標エリアに到達することができたが、単独学習の場合、2 体の個体の経路はより長かった。

なお、SGA 学習法の方策改善は、報酬を用いた内部変数（パラメータ）の勾配（付録 C.2 の Characteristic Eligibility）によって行うため、報酬値の設定は学習の収束に大きく影響する。2 個体シミュレーションにおいて、壁、障害物に衝突する場合は-1.0、ゴールエリアに到達する場合は+3.0 と設定し、経験上の最も良い収束を得た（群学習の場合は、両個体のユークリッド距離が 1.0～3.5 以内であれば+1.0、それ以外は-1.0 の報酬を与えた）。各内部変数の初期値は $\mu \in (0,1)$の乱数、p=0.5、β=1.0 で、学習率はそれぞれ、0.1、0.1、0.001 であった。減衰率 γ=0.997、D_i 初期値 0.0、b=0.6 とし、行動決定の閾値 θ_1= -1.5、θ_2= +1.5 とした。

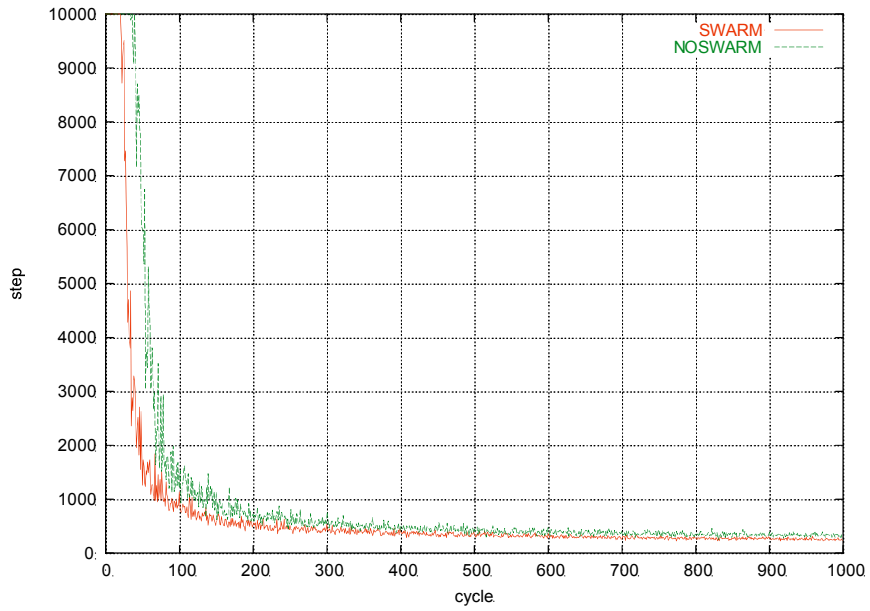
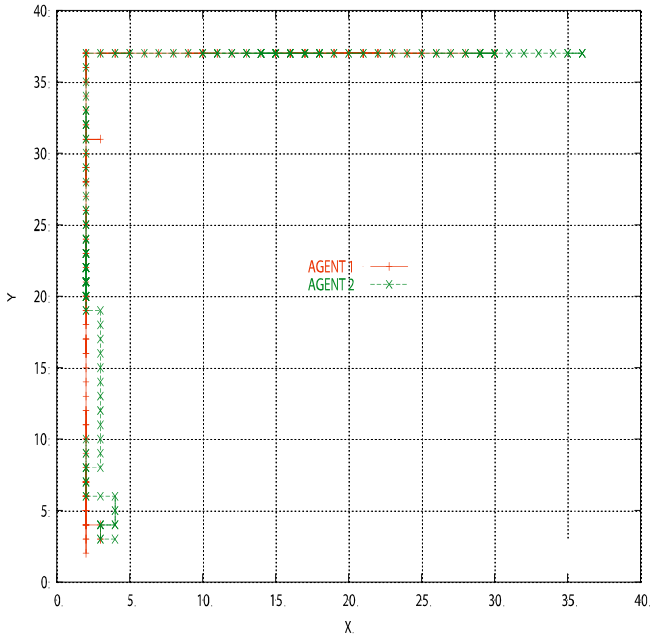
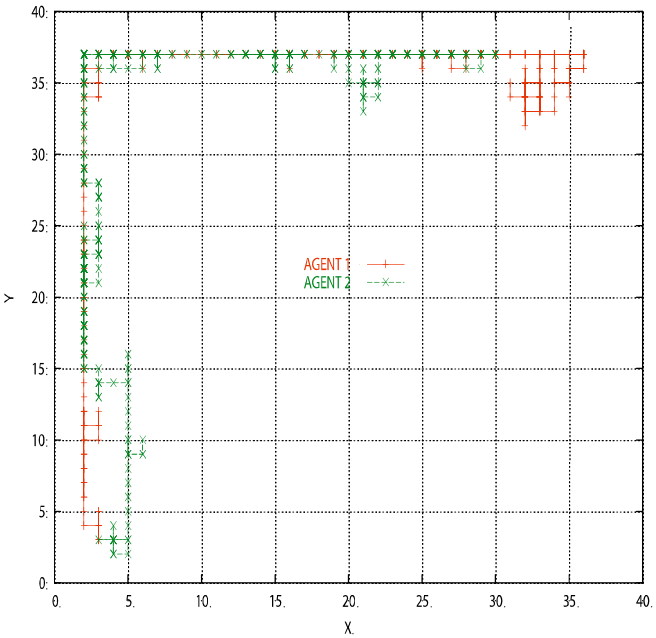


図 4.7 従来法 3: SGA 学習を用いた探索における経路長の変化



(a) 群学習の場合



(b) 単独学習の場合

図 4.8 従来法 3: SGA 学習を用いた学習終了時の探索軌跡

4.1.5 提案法 FAC の場合

第 3 章で提案したファジィネット(FN)を用いた Actor-Critic 学習型強化学習システム(FAC)の探索シミュレーションは状態遷移過程がマルコフ決定過程（MDP）の場合（3.1 節）と、状態の不完全知覚問題(incomplete perception problem または aliasing problem)が存在する部分

観測マルコフ決定過程（POMDP）の場合（3.2 節と 3.3 節）について、それぞれ行った。

POMDP の場合、学習による探索経路長の変化は、図 3.15 で示され、単独学習(Individual Learning)場合と群学習（Swarm Learning）の場合共に試行回数（cycle）の増加とともに減少し、収束が見られた。しかし、収束時の平均経路長は、個体の位置座標情報を入力とするMDP 場合(図 3.5)と異なり、いずれも最適解の 56step、または準最適解の 62step から大幅に離れ、群学習の場合は 747.9step、群行動を考慮しない単独学習の場合は 2,689.4step であった。

4.1.6 提案法 FQ の場合

第 3 章で提案したファジィネット(FN)を用いた Q 学習型強化学習システムの探索シミュレーションは既に 3.2 節で述べた。

学習による探索経路長の減少及び収束は、図 3.21 で示され、単独学習(Individual Learning)場合と群学習（Swarm Learning）の場合共に最適解の 56 step に達しなかったが、準最適解の 62 step に収束した。

なお、1,000 回試行で探索した経路長の 10 回のシミュレーションの平均値は、群行動を考慮した群学習の場合は 117.6 step/cycle、群行動を考慮しない単独学習の場合は 611.2 step/cycle で、群学習の方が優れた学習性能を持つことが確認できた。

学習終了時の両個体の探索経路は、図 3.22 で示された。群学習（図 3.22(a)）と単独学習（図 3.22(b)）の経路長はいずれも 62step のマンハッタン距離であったが、群学習の場合は、両個体が常に同じ経路を通ることに対し、単独学習の場合は両個体の経路が異なることになっている。

4.1.7 提案法 FS の場合

第 3 章で提案したファジィネット(FN)を用いた Sarsa 学習型強化学習システムの探索シミュレーションは既に 3.3 節で述べた。

学習による探索経路長の減少及び収束は、図 3.26 で示され、単独学習(Individual Learning)場合と群学習（Swarm Learning）の場合共に FQ と同様に、準最適解の 62step に収束した。

1,000 回試行で探索した経路長の 10 回のシミュレーションの平均値は、群行動を考慮した群学習の場合は 97.6 step/cycle、群行動を考慮しない単独学習の場合は 885.4 step/cycle で、群学習の方が優れた学習性能を持つことが確認できた。また、群学習の場合は、単独学習の場合に比べ、より迅速に収束したことが確認できた。更に、図 3.21 に見られた FQ 学習の収束後の振動現象が、FS の学習性能を示す図 3.26 では認められなくなった。

学習終了時の両個体の探索経路は、図 3.27 で示される。群学習（図 3.27(a)）と単独学習（図 3.27(b)）の経路長はいずれも 62step のマンハッタン距離であったが、群学習の場合は、両個体が常に同じ経路を通ることに対し、単独学習の場合は両個体の経路が異なっている。

表 4.1　各手法の学習性能の比較（POMDP 下の目標探索シミュレーション結果）.

学習方式	群学習有無	学習コスト (step/cycle)	収束時刻 (step No.)	終了時経路長 (step)
ランダム探索（学習なし）	群学習	5,788.9	該当なし	9,999.0
	単独学習			
Q 学習（従来法 1）[28]	群学習	3,147.4	390.0 （不安定）	3,048.0
	単独学習	6,041.1	320.0	6,024.0
Sarsa 学習（従来法 2）[29]	群学習	2,450.1	306.0	2,048.0
	単独学習	2,496.4	380.0	2,048.0
SGA 学習（従来法 3） [46]	群学習	733.2	不安定	232.0
	単独学習	1,045.5	不安定	327.0
FAC（提案法）[85]-[88] [94]	群学習	-	不安定	747.9
	単独学習	-	不安定	2,689.4
FQ （提案法) [92] [94]	群学習	117.6	93.0	62.0
	単独学習	611.2	333.0	62.0
FS （提案法） [92]-[94]	群学習	97.6	28.0	62.0
	単独学習	885.4	407.0	62.0

注１：最短経路長（理論値）は 56 step　である。

注２：数値は 10 回シミュレーションで各シミュレーションにおいて、試行回数(cycle)が 1000 回とする場合の平均値である。

4.1.8 各手法のシミュレーション結果の比較

POMDP 環境での目標探索問題に対し、従来の強化学習手法（Q 学習、Sarsa 学習、SGA 学習）と本論文で提案したファジィネットを用いた強化学習システム（FAC、FQ、FS）の学習性能の比較を、これまで述べたシミュレーション結果により、表 4.1 及び図 4.9、図 4.10 にまとめる。

表 4.1 及び図 4.7 より、従来法の Q 学習[15][28][31]、Sarsa 学習[15][29][31]はランダム探索より短い経路長を発見することができたが、Sarsa の単独学習の平均経路長より長くなる FAC の単独学習の場合を除いて、提案法の FAC [85]-[88] [94]、FQ [92] [94]、FS [92]-[94]に比べ、いずれも従来法は劣っていることが分かる。また、従来法の SGA 学習[46]は単独学習及び群学習のいずれの場合において提案法 FAC より短い経路(327step と 232step)を探索することができたが、提案法 FQ と FS の準最適解(いずれも 62step)に比べ、劣っていることが分かる。

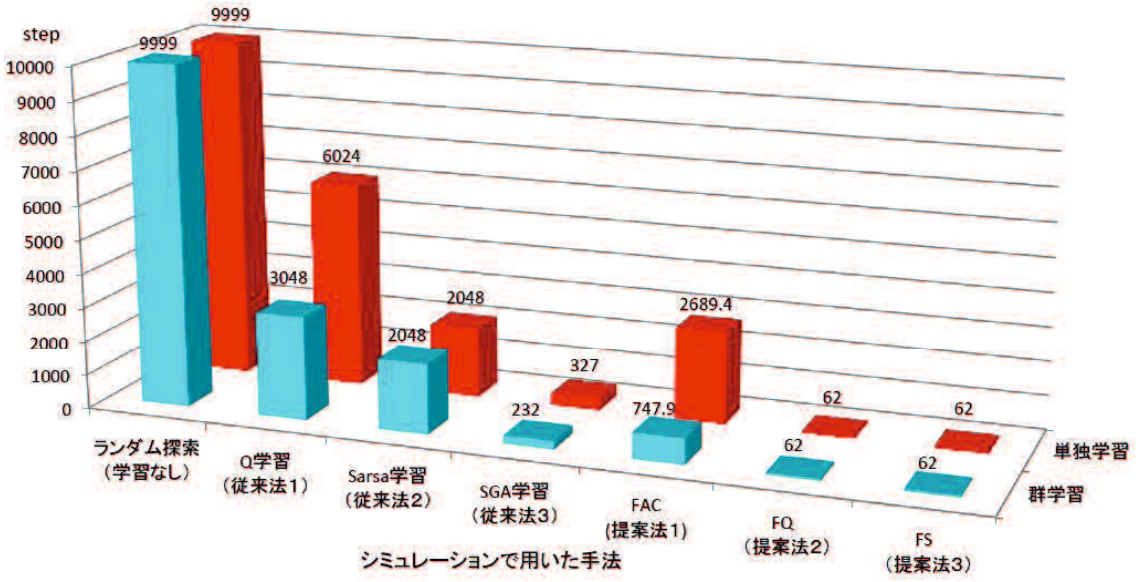
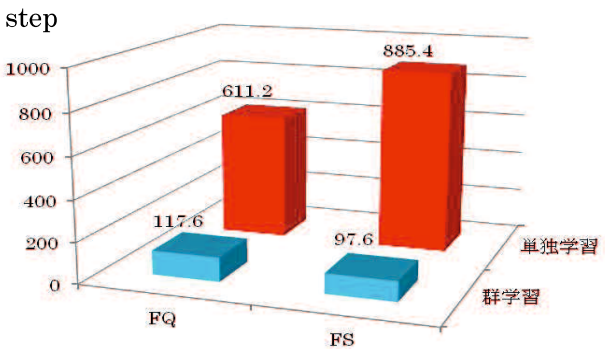


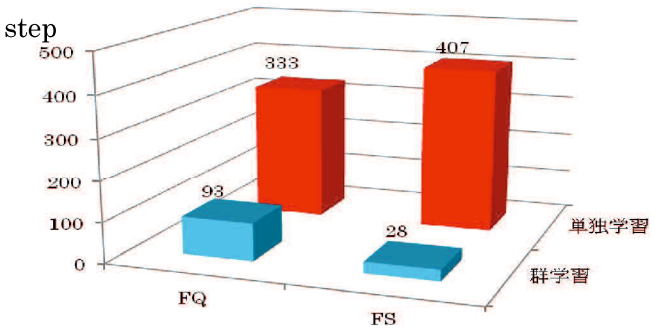
図 4.9 各手法による目標到達時の平均経路長

なお、各手法の収束後の経路長について、提案法の FQ と FS 共に準最短路 62 step に到達している。また、FQ と FS の学習収束の速度及び 1,000 試行(cycle)の平均経路長の比較は図 4.10 に示す。

図 4.10(a)の平均経路長と図 4.10(b)の収束速度について、いずれも群学習の場合は FS が優位に、単独学習の場合は FQ が優位に立つことが明らかになった。現状態と次状態の二つの状態―行動価値関数を用いる Sarsa 学習を用いた FN 型強化学習システムが個体間の距離を考慮した群学習により適していることが考えられる。



(a)　平均経路長(step/cycle)



(b) 平均収束時刻(step No.)

図 4.10 提案法 FQ と提案法 FS の学習性能の比較

4.1.9 各手法のシミュレーションの計算時間

表 4.1 にリストされたランダム探索、従来の強化学習法である Q 学習、Sarsa 学習、SGA 学習及び本論文で提案した強化学習システム FAC、FQ、FS を用いた部分観測マルコフ決定過程（POMDP）の未知環境探索シミュレーションの計算時間は計算機ハードウェアやプログラムの構成によるものであるが、（株）デルのワークステーション Precision 360（CPU: 3.2 GHz、メモリ：1GB）の場合、1 回のシミュレーション時間は、数秒―数分程度であった。強化学習の確率的探索の特徴によって、収束する場合としない場合の探索シミュレーションの計算時間は大幅に異なる。よって、ここではシミュレーションの計算時間の詳細な考察は省き、前節まで示した計算コストなどの学習性能によって、各手法を比較した。

4.2 方策関数におけるパラメータの設定

第 2 章で述べた行動方策には、Max-Q、*ε-greedy* 及び Soft-Max 法があるが、第 3 章で提案した FAC、FQ、FS の場合はいずれも探索(exploration)と知識利用(exploitation)を同時に行うことが可能な Soft-Max 法を用いた。FAC は(3.3)式、FQ は(3.10)式、FS は(3.20)と(3.21)式を用いて、行動価値関数(FAC の場合)A、または、状態―行動価値関数(FQ と FS の場合)Q の値に応じたボルツマン確率分布に従い、行動選択を行った。これらの確率方策関数には「温度定数」（*T*）というパラメータがあり、温度 *T* が大きいほど価値関数の値への依存が小さく、よりランダム探索するように行動が選択される。

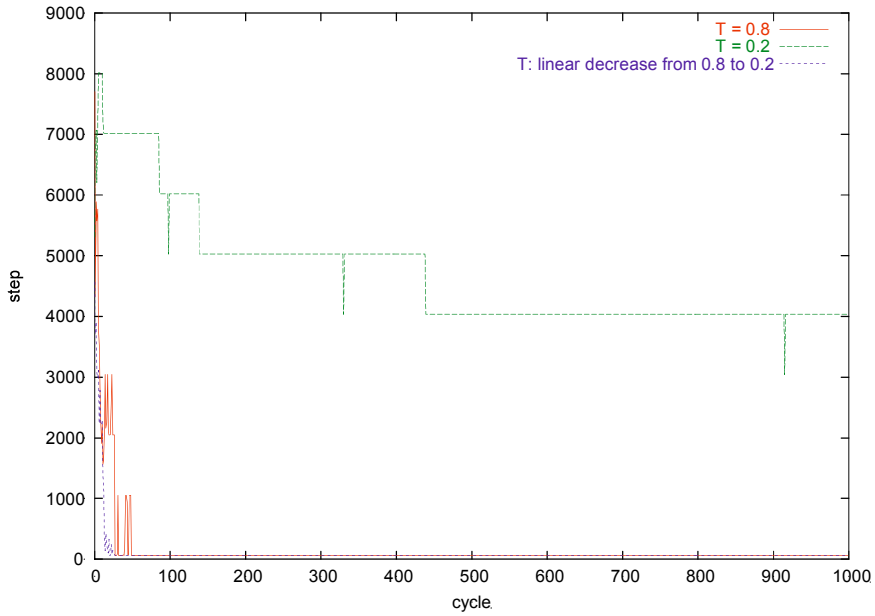


図 4.11 温度定数 *T* が固定値の場合（*T*=0.2, *T*=0.8）と線形的減少の値を用いる場合の学習性能の比較

提案法の各シミュレーションにおいて、学習の収束様子を観測しながら経験的に温度定数を決定した。ここでは、第 3.3 節の FS の場合を例として、温度 *T* の経験上の決定法を示す。

まずは、*T* が固定の値（*T*=0.8 と *T*=0.2）の場合と「探索した経路長の収束に従って線形的に減少する」場合の FS の学習グラフを図 4.11 に示す。図 4.11 では、*T*=0.2 の場合は状態の探索は不十分のままで目標までの経路長が収束せず、4,000 step 付近で局所解に陥った。一方、*T*=0.8 の場合は、準最短経路 62 step に収束するものの、初期の学習コスト（step 数）が高く、温度 *T* が「linear decrease from 0.8 to 0.2」の場合と比較し、性能が劣っている。なお、*T* の値を線形的に減少させるタイミングは経路長の収束の見られる 200 cycles から始まり、計算法は(4.1)式に示す。

$$T \leftarrow T - 0.003 * cycle + 1.4 \tag{4.1}$$

また、経路長の収束に従って T 値の減少は線形でなく、非線形的な場合は、学習の収束がより迅速に実現することが実験によって分かった。図 4.12 は指数関数を用いた(4.2)式によって決定した温度定数を使用した場合（実線）と線形的減少の場合（破線）の収束を示している。

$$T = 0.8/(1 + \exp(0.03 * cycle - 5) + 0.2 \tag{4.2}$$

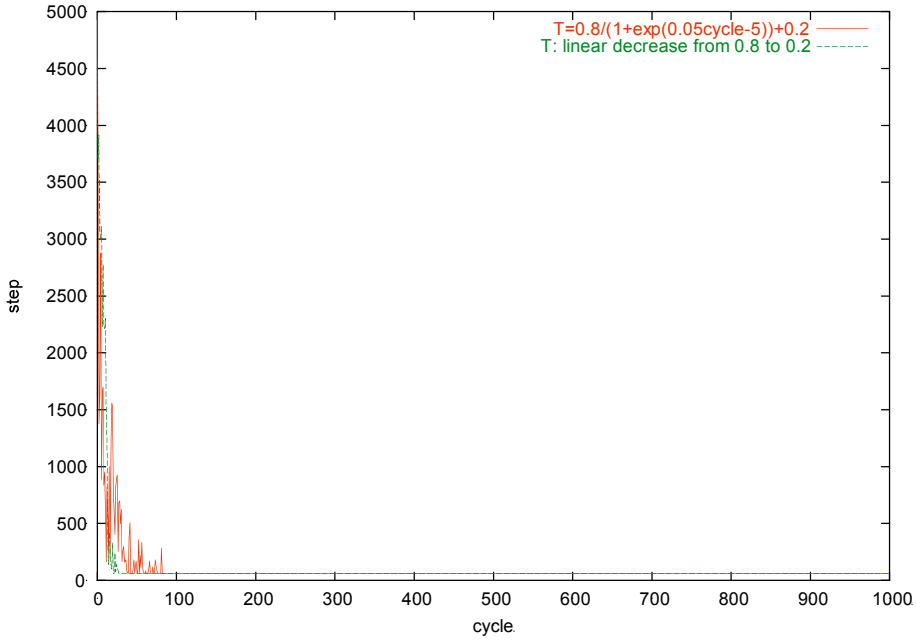


図 4.12 温度定数 *T* が線形的減少と非線形的減少の場合の学習性能の比較

なお、(4.1)式の線形的減少関数及び(4.2)式の指数的減少関数の形を図 4.13 に示す。本論文 3.3 節での FS の探索シミュレーションにおいて、経験上最も性能の良い線形的減少関数の温度定数を使用していたが、今後更なる理論的な検討が必要であろう。

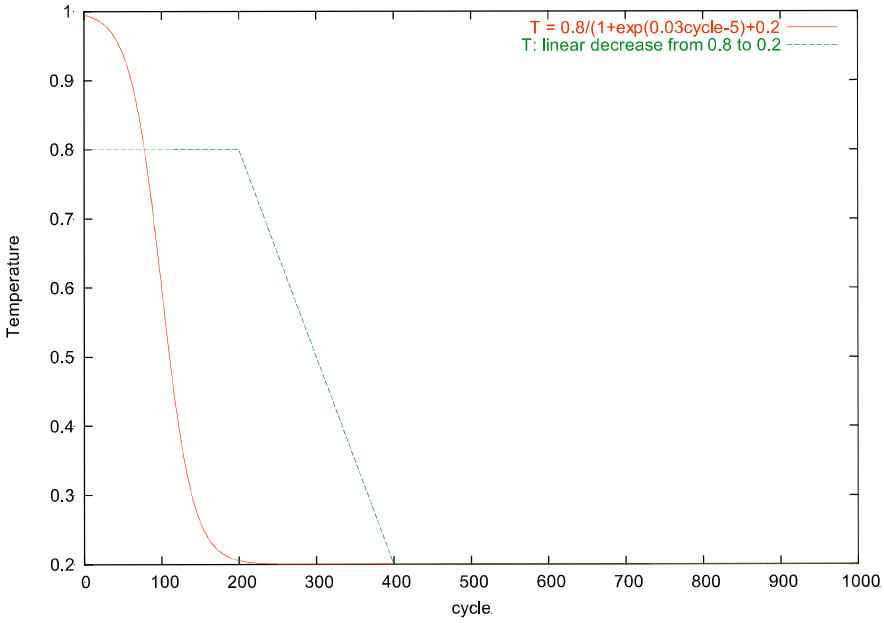


図 4.13 試行（学習）回数と共に非線形的、または線形的に減少する温度定数 *T*

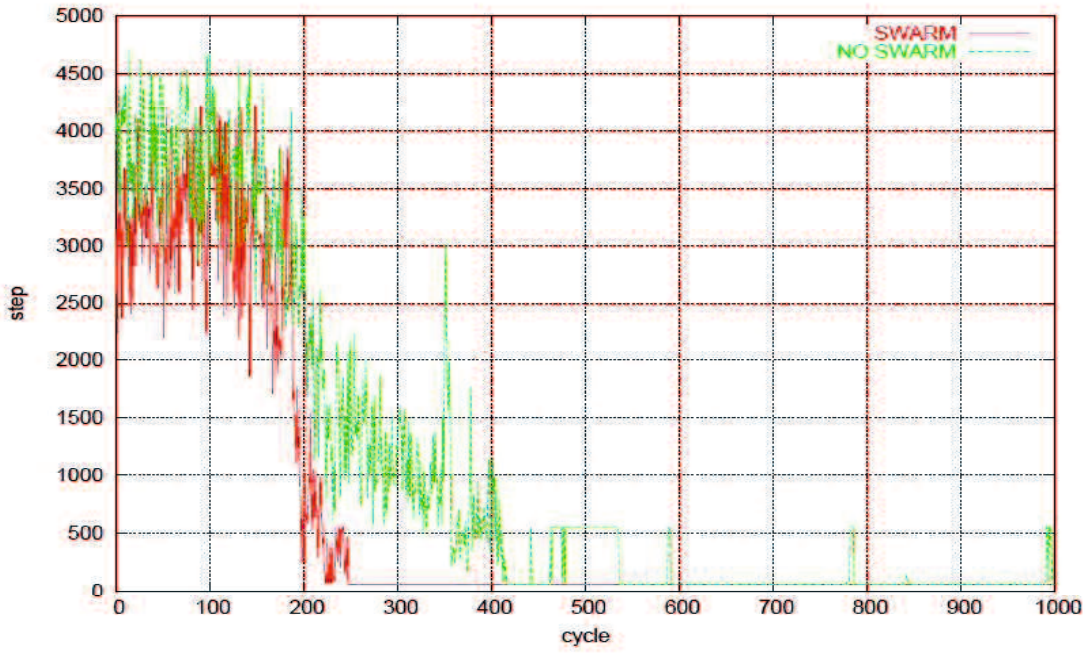
4.3 学習率の設定

ニューラルネットワークの学習は、ユニット（ニューロン）間の結合荷重を修正することによって実現する。修正の幅を決めるパラメータは「学習率」と呼ばれる。一般的に、学習の収束を促進し、出力の振動を抑制するため、学習回数の増加につれ、学習率を減衰させることが有効である。本論文で提案した FAC を除き、FQ の学習則(3.1)式、FS の学習則(3.15)式、(3.16)式の学習率が、固定の値ではなく、Derhami ら[20]が提案した適応学習率(ALR: Adaptive Learning Rate）を導入している。ここで、FS の場合について、固定の学習率（α＝0.3）と学習回数や状態の訪問回数に応じて変化する適応学習率((3.15)式、(3.16)式)をそれぞれ用いる場合の強化学習システムの学習性能について、シミュレーションの結果によって比較する。

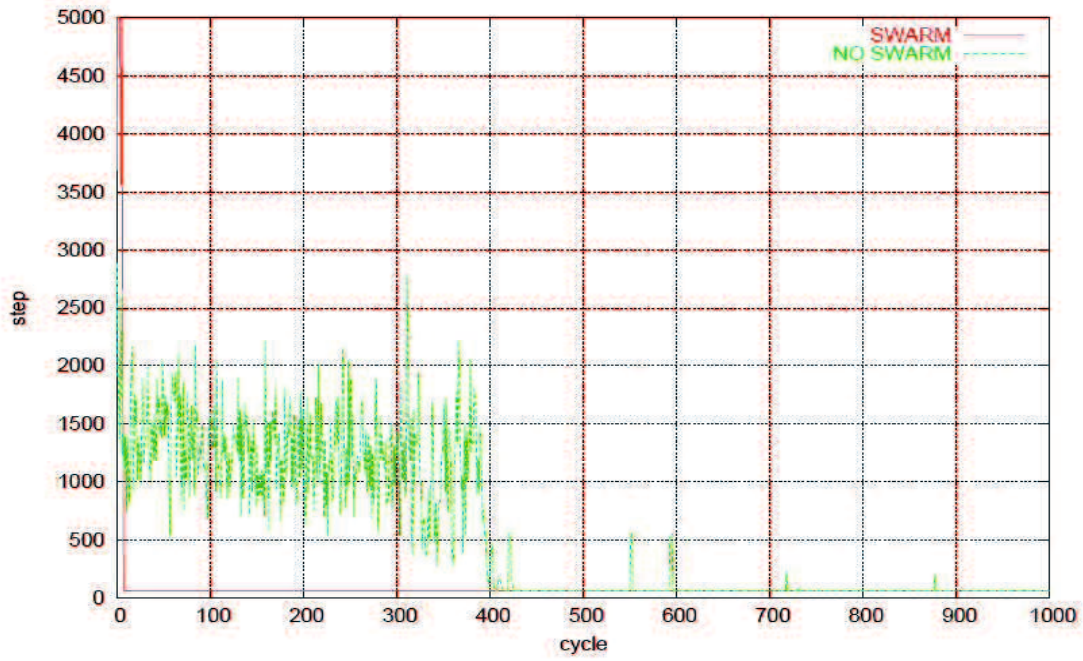
障害物がない POMDP 環境（すなわち、個体の位置座標でなく、近傍の観測情報を入力とする）（図 4.3）と、障害物ありの POMDP 環境（図 3.20）を用いて、2 体の個体が FS を用いた目標探索シミュレーションを行った。

図 4.14 は障害物がない環境の場合の学習率の設定法によって FS の学習性能の違いを示している。群学習（SWARM）と単独学習（NO SWARM）の場合共に適応学習率（ALR）の方が収束早いことが分かる。特に群学習の場合は 0.3 に固定した学習率（図 4.14(a)）より、ALR 法(図 4.14(b))の有効性が明らかになっている。

類似した結果が障害物ありの環境でも得られた。図 4.15 より、ALR 法の有効性が確認される。

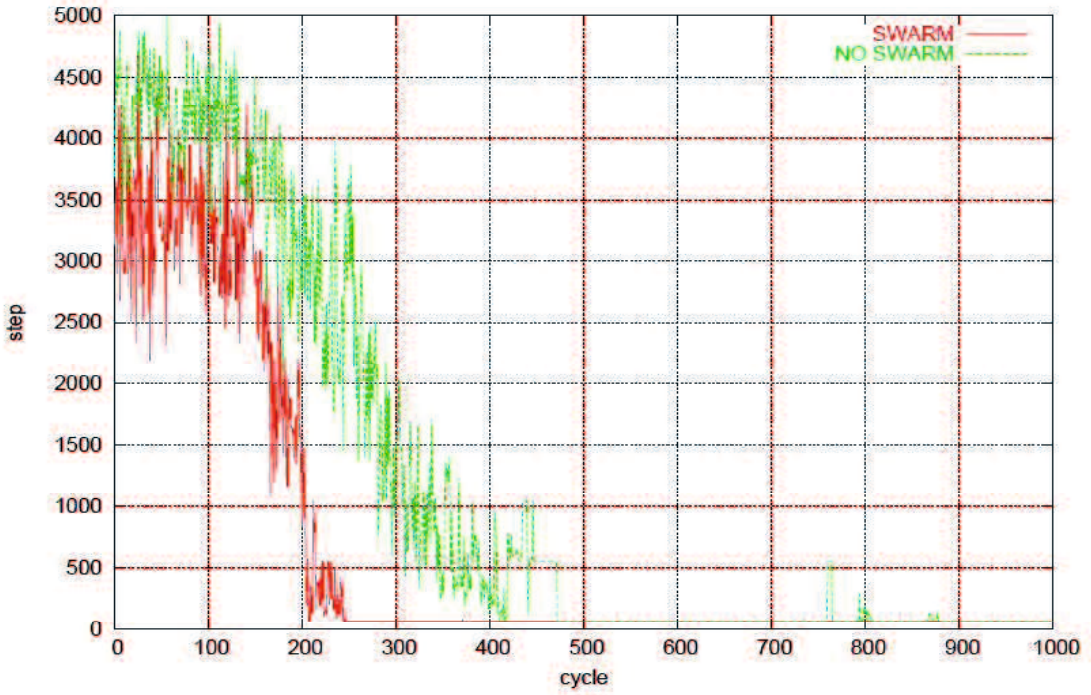


(a) 学習率を $\alpha = 0.3$ に固定した場合（FS）

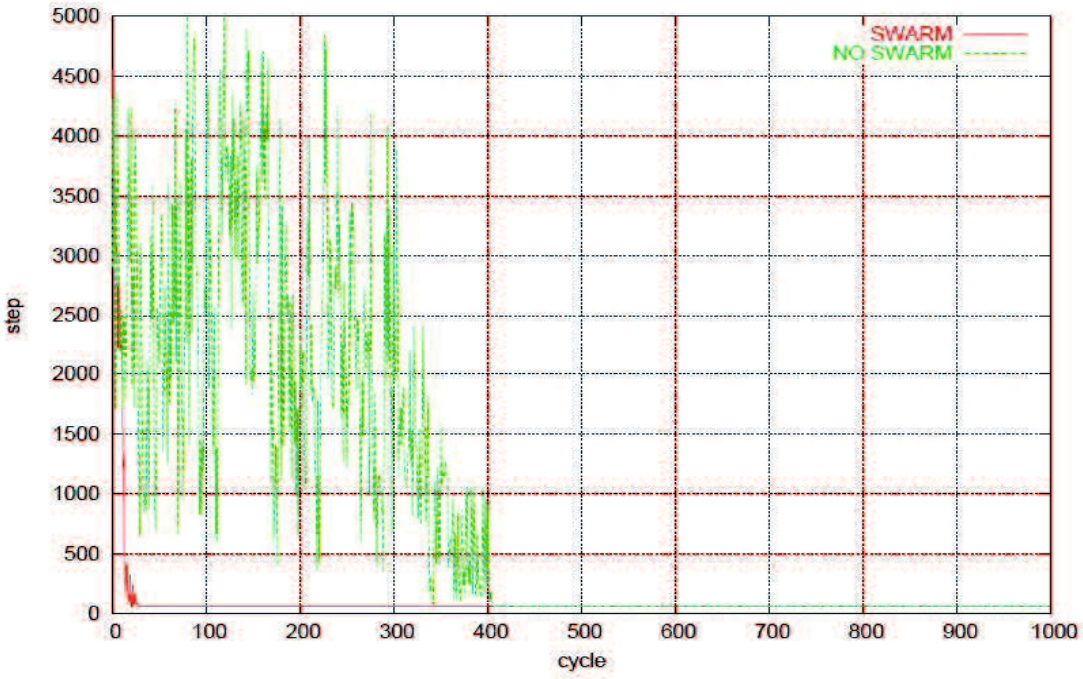


(b)学習率 α^{ALR} を 0.001～0.3 の間に変動させる場合（FS）

図 4.14 学習率による学習性能への影響（障害物なし（図 3.3）の場合）



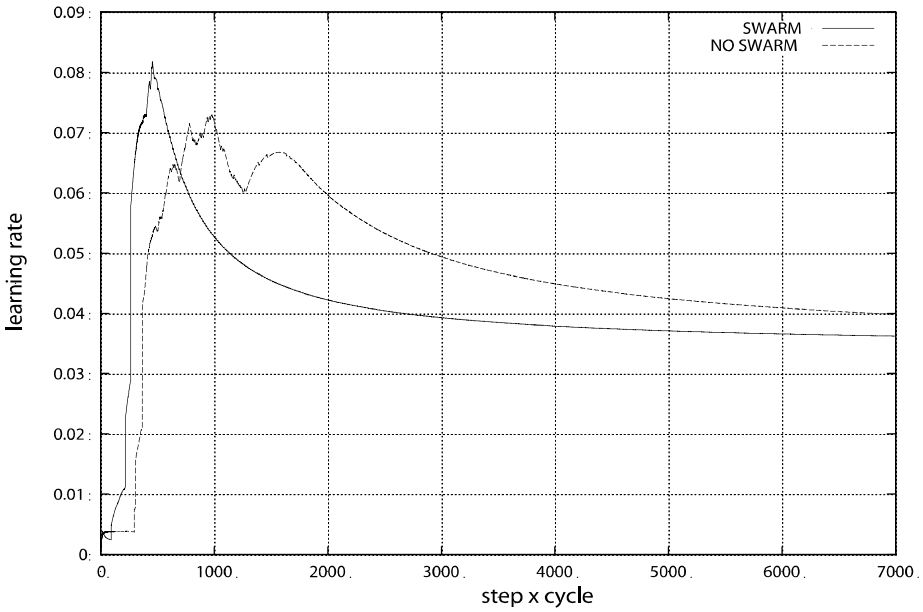
(a)　学習率を $\alpha = 0.3$ に固定した場合（FS）



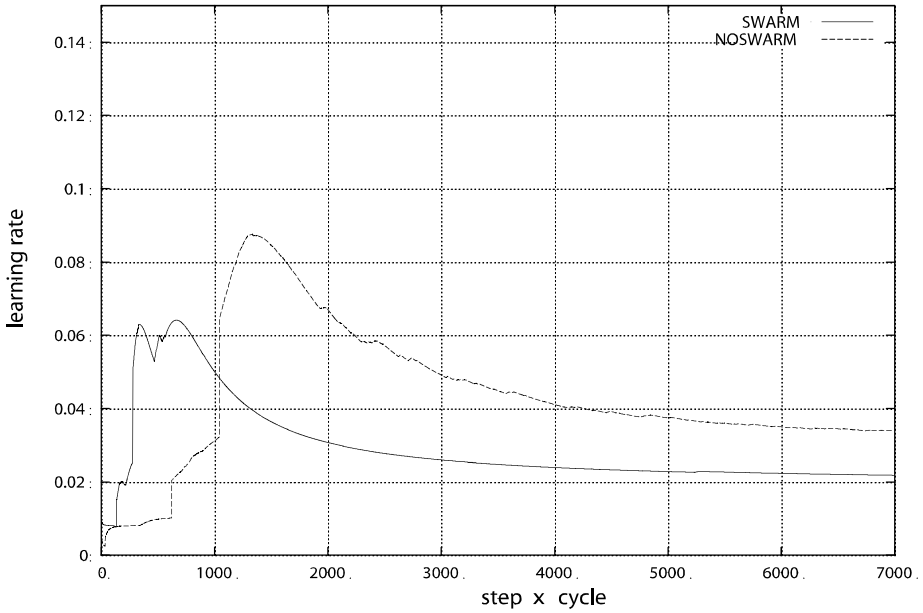
(b)学習率 α^{ALR} を 0.001～0.3 の間に変動させる場合（FS）

図 4.15 学習率による学習性能への影響（障害物あり（図 3.20）の場合）

図 4.16 は ALR 法を使用した場合，ある状態に関する適応可能な学習率 α_t の変化を示す（具体的には、個体 a_1 の状態「0000」、すなわち上下左右がすべて通路の場合である）。探索 step および試行回数 cycle ごとに α_t が更新されるため、横軸は更新回数の step×cycle となっている（更新回数 7,000 以降の収束カーブを略す）。初期の探索段階においては、該当状態の訪問回数が少なく、学習率 α_t が増加したが、学習の繰り返しによって他状態への訪問が減少し、「通路 0000」状態への訪問回数が増え、適応学習率が減少しながら収束することになった。なお、いずれも群学習の場合の ALR の収束が早く、収束時の値が低かった。



(a) 障害物なし（図 3.3）POMDP 環境の場合



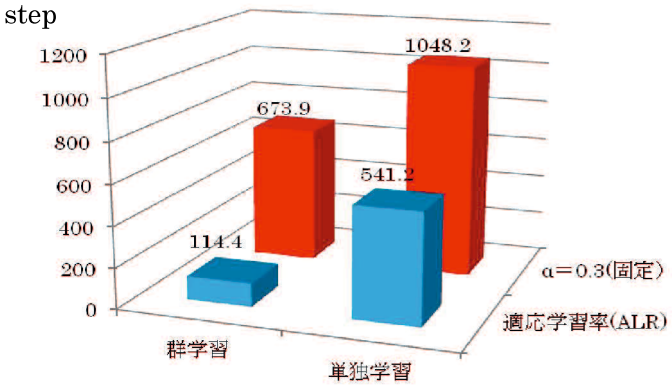
(b) 障害物あり（図 3.20）POMDP 環境の場合

図 4.16 学習過程における適応学習率(ALR)の変化(FS)

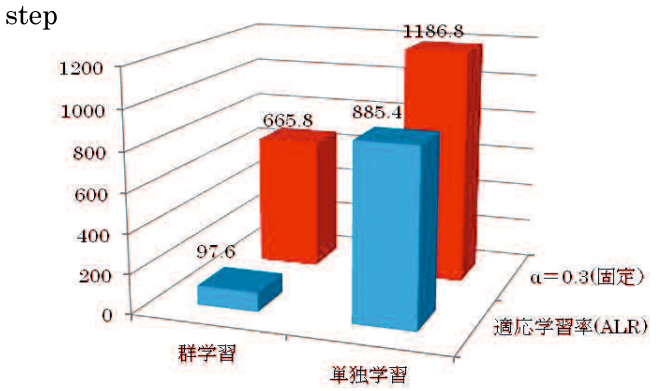
なお、ALR が収束した時の値や、固定値の学習率との学習性能の詳細の比較を表 4.2 に示す。1,000 試行（学習）の平均経路長の比較を図 4.17 に示している。同図より、ALR 及び群学習の有効性がより明白に確認できる。

表 4.2　FS における学習率設定による学習性能の比較(1,000 試行平均)

探索環境	群学習有無	ALR(α_t)収束値*	α_t=ALR で平均経路長	α_t= 0.3 で平均経路長
障害物なし（図 3.20）	群学習	0.0362	114.4	673.9
	単独学習	0.0398	541.2	1,048.2
障害物あり（図 3.20）	群学習	0.0217	97.6	665.8
	単独学習	0.0339	885.4	1,186.8



(a)　障害物なし（図 3.20）POMDP 環境の場合



(b)　障害物あり（図 3.20）POMDP 環境の場合

図 4.17 FS における異なる学習率による学習コストの比較（1,000 試行の学習時の平均経路長）

第 5 章

まとめと今後の課題

本論文では、未知環境に適応できる知的個体の内部モデルを提案した。提案したモデルは自己組織化ファジィニューラルネットワーク（SOFNN）と強化学習システムによって構成された。前者は、システムへの入力情報を分類・識別するために設計され、これまで、カオス時系列などの非線形予測[38][39][41][42]、インテリジェント制御[43]-[46]分野でもよく使用されている。後者の強化学習システムは、従来の Actor-Critic 学習[15][23]-[27]、Q 学習[15][28][31]、Sarsa 学習[15][29][31]を、それぞれ用いて、SOFNN との結合によって、「FAC」、「FQ」と「FS」といった異なる強化学習システムを含む。これらの強化学習システムは、それぞれ、以下の特徴を持つ(表 5.1 参照)。

(i) SOFNN を用いた Actor-Critic 型強化学習システム FAC の場合は、状態価値関数（Critic）の変化である時間差分誤差（TD 誤差）より、適切な行動を選択するように行動価値関数（Actor）を修正し、確率方策関数の学習を行う。行動選択と状態評価を同時に行われるため、政策オンの学習が MDP 環境では実現できるが、POMDP 環境では最適解への接近は困難である。

(ii) SOFNN を用いた Q 学習型強化学習システム FQ の場合は、次状態の最大値を持つ状態―行動価値関数（Q 関数）を直接 TD 誤差に導入し、方策オフの学習アルゴリズムによって、現状態の Q 値を修正し、確率方策関数の学習を行う。常に蓄積した知識を最大限に利用するため、POMDP 環境で準最適解への収束が迅速であるが、その反面安定性に欠ける。

(iii) SOFNN を用いた Sarsa 学習型強化学習システム FS の場合は、現状態と次状態の状態―行動価値関数（Q 関数）をそれぞれ TD 誤差に導入し、方策オンの学習アルゴリズムによって、現状態の Q 値を修正し、確率方策関数の学習を行う。FQ に比べ、より多くの状態―行動価値関数を使用するため、POMDP 環境で準最適解への収束が安定であるが、その反面学習コストはやや高くなる。

また、複数の個体が互いの位置と距離を観測し、鳥や魚などの群れを形成するような BOID ルール[110][111]を強化学習の報酬に導入した。個体間の適切な距離を保持する「群学

習」の場合は、粒子群最適化（PSO）[112][113]やアントコロニー最適化(ACO)などの「群知能」手法の特徴である最適解への多点探索ができるため、個体ごとに、それぞれ独立に環境・目標を探索する「単独学習」より、システム全体の学習性能が優れている。POMDP 環境での目標探索シミュレーションにおいては、提案法の中でも、群学習を用いた FS の学習性能が最も優れ、推奨手法である（表 5.1）。

表 5.1　提案手法間の比較: POMDP の場合

提案手法	価値関数	最適解への収束	収束速度	収束安定性
FAC	行動	非最適解	最も遅い	最も不安定（図 3.15）
FQ	状態・行動対	準最適解	単独学習：早い 群学習：遅い	不安定（図 3.21）
FS	状態・行動対 （連続 2 回）	準最適解	単独学習：遅い 群学習：早い	安定（図 3.26）

本論文で提案した強化学習システムは、簡単な計算機シミュレーションによってその有効性が確認できたが、今後、未知環境で適応行動を獲得するための自律ロボットの内部モデルとして応用されることが期待できる。その理由は以下に挙げられる。

(i) 提案システムには、多次元入力情報に対する自動的な分類・識別機能を持ち、外部環境の知識を獲得することができる。例えば、実機ロボットに応用する場合は、環境情報を、画像や音声などの信号によって取り込み、提案システムのファジィネット(FN)で自動的に分類または識別し、強化学習に存在する状態数の爆発問題を解決することができる。

(ii) 提案システムには、確率的であるが、能動的な出力を行うことができ、外部環境に適応する行動を獲得することができる。例えば、実環境における自律ロボットの行動制御は、事前に設計したプロセスでなく、ロボット自身の試行錯誤によって学習することができる。

(iii) 提案システムには、他個体との距離を適切に取ることにより、群探索することができる。従って、システムの全体の学習性能の向上を実現することができる。例えば、深海や宇宙空間などの未知環境探索や、災害救助（被害者捜査など）の場合、多数の自律移動ロボットによって行うことが効率よく、本論文で提案した群学習手法が応用できる。

なお、複数の個体の適応・協調行動は、本論文では互いの距離に正、または負の報酬を与えることのみで、ある程度実現した。今後、好奇心、情動や感情などより高度な認知・心理モデル[128]-[132]の導入によって、個体群の創発(emergence)を含め、更に改良を行っていきたい。

謝 辞

平成 5 年（1993 年）4 月から、筆者が山口大学にて、人工神経回路網（ANN: Artificial Neural Network）を始め、動画像処理、機械学習(Machine Learning)、カオス時系列予測などの勉学と研究を行ってきた。本論文は、筆者がこれまで行ってきた研究の中で、ファジィニューラルネットワークを用いた強化学習システムに関する研究をまとめたのです。

本研究は平成 14 年（2002 年）から始まり、約十年間を費やしてある程度の成果を上げたものです。まず、申し上げなければならないのは、自己組織化ファジィニューラルネットワークの考案は、山口大学大学院理工学研究科情報・デザイン工学系学域大林正直教授によるものです。また、知的エージェントや自律ロボットの今後の発展を支える重要な技術とされる強化学習研究への興味も大林先生の適切なご助言とご指導をいただいたものです。さらに、この間、日常の勉学・研究環境を整えていただき、本論文をまとめるに当たり、大林先生のご支援がなけらば、筆者が到底できないものだと思います。ここに心から深く感謝を申し上げます。

また、山口大学大学院理工学研究科石川昌明教授、同田中幹也教授、同浜本義彦教授、松藤信哉教授には、本論文の完成に当たり、適切なご教示ご助言を頂きました。ここに深く感謝を申し上げます。

山口大学大学院教育学研究科古賀和利教授には、筆者が山口大学大学院博士後期課程在学中の指導教員として、勉学・研究環境を整えて頂いた上、「常に原点に戻る」と言われる学問や研究に対する取るべき姿勢を教えて頂きました。また、外国人留学生であった筆者に対し、日常生活を含め、格別なご厚愛ご芳情を賜りまして頂きました。ここに心から深く感謝の意を申し上げます。

元山口大学工学部教授の鳥岡豊士（故人）には、筆者が山口大学大学院博士前期課程在学中の指導教員として、勉学・研究環境を整えて頂いた上、筆者を脳型情報処理分野の研究方向へ導いた最初の師であり、本研究の礎を築く学問を教えて頂きました。ここに心から深く感謝の意を申し上げます。

元山口大学大学院理工学研究科（現愛知県立大学情報科学部）小林邦和准教授には、筆者が山口大学の敷地に入る第 1 歩から本論文をまとめた本日まで、公私ともに格別なご友情ご支援を頂いた上、様々な有益なご助言ご教示を頂きました。ここに心から深く感謝の意を申し上げます。

山口大学工学部知能情報工学科（旧知能情報システム工学科）生体情報システム工学研究室（旧生体情報工学研究室)、並びに同情報機器学研究室に在籍された卒業生・修了生には、日頃の勉学・研究に当たり、真摯なご友情ご支援を頂きました。特に現山口大学大学院理工学研究科水上嘉樹准教授、現広島工業大学生命学部塚本壮輔准教授、現富士通（株）の足立浩一氏、現日本電気（株）の米田賢太郎氏、現富士通テン（株）の山野祐樹氏、並びに、現山口県庁の梅迫公輔氏、現大分シーイーシー（株）の江藤剛士氏、現安川情報システム（株）の山本歩氏、北野寛明氏（故人)、現富士通（株）の大宮理恵氏、現三菱電機インフォメーションシステムズ（株）の小川長久氏、現富士通（株）の羽野ともえ氏、寺森夏樹氏、現（株）メイテックの杉野元紀氏、現富士通（株）の大田智範氏、現（株）西日本情報システムの松崎洋一郎氏、現日本電気（株）の古本隆人氏、現九州日本電気ソフトウェア（株）の波多 聡氏、現（株）ハイエレコンの木下康弘氏、現東京エレクトロン九州（株）の木村慎祐氏、現（株）アイ・エル・シーの山根哲也氏、現（株）ハイビスタイルの真境名 郁氏、現中国科学院深せん先進技術研究院の馮 良炳氏、現（株）宇部情報システムの鶴崎徹也氏などの方々に感謝の意を表します。

また、私の小中高及び大学時代の先生方、同窓の諸君に感謝の意を表します。

中国から日本への留学に当たり、元長野県飯田姫長高等学校の望月恆朗先生、同山下守弘先生には暖かいご支援を頂き、大変お世話になりました。また、母の友人である篠田啓子様を始め、飯田市の親切な人々、元日本歯科大学教授の青木茂治先生には、私の留学のご支援を頂きました。東京都新宿区にある国際学友会日本語学校（現日本学生支援機構東京日本語教育センター）の先生方、東京都新宿区神楽坂にある橋爪商店の皆様、同窓の諸国の皆さんには、忘れられない日本での最初に一年を励まして頂きました。ここに心から深く感謝の意を申し上げます。

永山克昭・良子ご夫婦様、大久保義雄・和子ご夫婦様を始め、白石勝己様、安光紀晶様、竹田康治様、高杉紀雄・静江ご夫婦様、山口大学留学生や外国人を大変ご熱心に世話して頂きました宇部市の多くの市民ボランティアの皆様には、私と私の家族の宇部市での生活を末永くご支援していただきました。ここに心から深く感謝致します。

英国マンチェスター大学計算機科学スクール准教授の Ke Chen 博士には共同研究を通して、多くの有益な助言を頂き、筆者の研究が深められました。特に平成 20 年 6 月～7 月に筆者を同大学の客員研究員としてお招き頂き、筆者に機械学習などの情報学の最先端の研究方向をご教示頂きました。ここに心から深く感謝の意を申し上げます。

元宇部フロンティア大学短期大学部情報システム学科教授の吉田信夫先生、元広島工業大学工学部電子情報工学科教授の原 肇先生には、活発な研究討論を通して、筆者の研究が推進されました。ここに心から深い謝意を申し上げます。

電子情報通信学会(IEICE)、計測自動制御学会(SICE)、電気学会(IEEJ)、米国電気電子学会(IEEE)、国際予測学会(IIF)など主催の国際会議、研究会、シンポジウム、全国大会、支部大会に参加された多くの研究者の方々には、有益な助言を頂いており、筆者の研究が推進されました。

また、本研究の成果を公開させていただいた知的計算国際会議(ICIC 2005～ICIC 2013)、IEEE 計算知能世界大会（WCCI/IJCNN 2008）、Journal of Circuits, Systems, and Computers (JCSC)、International Journal of Intelligenct Computing and Cybernetics (IJICC)、電気学会論文誌 C の多数の査読者には、適切な助言を頂きました。ここに深く感謝の意を表します。

本研究は山口大学の運営費交付金を始め、多くの研究費支援のお陰で行うことができました。

文部科学省、日本学術振興会には、以下の科学研究費補助金により、研究費の支援を受けました。

１．　基盤研究（B）　(研究分担者)　課題番号：13450176，2001～2003 年度
２．　若手研究（B）　(研究代表者)　課題番号：40294657，2003～2005 年度
３．　基盤研究（C）　(研究代表者)　課題番号：18500230，2006～2008 年度
４．　基盤研究（C）　(研究分担者)　課題番号：20500277，2008～2010 年度
５．　基盤研究（C）　(研究分担者)　課題番号：20500207，2008～2010 年度
６．　基盤研究（C）　(研究分担者)　課題番号：23500181，2011～2013 年度
７．　基盤研究（C）　(研究分担者)　課題番号：25330287，2013～2016 年度（予定）

科学技術融合振興財団には、調査研究助成（研究代表者，2012～2013 年度）を受けました。また、山口大学工学部には、21 プロジェクト助成（研究代表者，2003～2004 年度）を受けました。ここに感謝を申し上げます。

山口大学学長、山口大学大学院理工学研究科長、山口大学工学部長、山口大学工学部知能情報工学科長を始め、山口大学の教職員の皆様には、本研究の推進に多大なご支援とご理解を頂きました。ここに心から深く感謝の意を申し上げます。

最後になりましたが、これまで筆者を支えた家族の皆に感謝したと思います。

本論文を一昨年に他界にしました私の祖母陳 敬蘭様に捧げたいものです。共働きの両親より、子供時代の私の成長をよく見守って頂きました。ありがとうございました。

私を生み、育てきた両親である合肥工業大学教授の呉 報任先生、安徽医科大学主任技師の劉 茂雲先生に深く感謝の意を申し上げます。

一緒に成長してきた仲の良い兄弟である舜君、千絵子ちゃんに心から感謝致します。

また、常に傍で支えてくれた妻（永子）、幸福を感じさせてくれた長女（宗雪）と長男（総太郎）に感謝致します。

皆様のご多幸とご活躍をお祈りいたします。

2013 年 8 月初稿
2013 年 12 月修正稿
呉本　尭

参考文献

[1] 松原仁：「Deep Blueの勝利が人工知能にもたらすもの」，*人工知能学会誌*，Vol.12, No.5, pp.698-703, 1997.

[2] RoboCup, http://www.robocup.org

[3] 金淵培：「エージェント技術の現状と実用化」，*人工知能学会誌*, Vol.12, No.6, pp.850-860, 1997.

[4] 三上貞芳：「強化学習のマルチエージェント系への応用」，*人工知能学会誌*, Vol.12, No.6, pp.845-849, 1997.

[5] Sycara K.P., “Multi-agent systems”, *Artificial Intelligence Magazine*, Summer, pp.79-92B, 1998.

[6] 荒井幸代，宮崎和光，小林重信：「マルチエージェント強化学習の方法論—Q-learningとprofit sharingによる接近—」，*人工知能学会誌*，Vol.13, No.4, pp.609-618, 1998.

[7] 荒井幸代：「マルチエージェント強化学習—実用化に向けての課題・理論・諸技術との融合—」，*人工知能学会誌*，Vol.16, No.4, pp.461-481, 2001.

[8] 浅田稔, 國吉康夫：「ロボットインテリジェンス」，*岩波書店*，2006.

[9] 土井利忠，藤田雅博，下村秀樹：「身体を持つ知能—脳科学とロボティクスの共進化」，*シュプリンガー・ジャパン*，2006.

[10] 八谷大岳，杉山将：「強くなるロボティックス・ゲームプレイヤーの作り方—実践で学ぶ強化学習」，*マイコミ*，2008.

[11] 大倉和博：「スワーム：群れの創発的挙動生成」，*計測と制御*, Vol.52, No.3, pp.179-182, 2013.

[12] 上田完次：「創発とマルチエージェントシステム」，*培風館*，2007.

[13] 高玉圭樹：「マルチエージェント学習—相互作用の謎に迫る」，*コロナ社*，2003.

[14] Sutton R.S, and Barto A.G.: “Toward a modern theory of adaptive networks: Expectation and Prediction”, *Psychological Review*, Vol. 88, pp. 135-170, 1981.

[15] Sutton R.S., and Barto A. G. (三上貞芳，皆川雅章 共訳):「強化学習」，*森北出版*，1998.

[16] Bellman R.E.: Dynamic Programming, *Princeton University Press*, 1957.

[17] Bellman R.E.: A Markov decision process, *Journal of Mathematical Mechanics*, Vol. 6, pp. 679-684, 1957.

[18] 牧野貴樹，澁谷長史：「強化学習最近の発展—総論：リレー解説「強化学習の最近の発展」にあたって」，*計測と制御*，Vol.52, No.1, pp.64-67, 2013.

[19] 木村元：「強化学習最近の発展—総論：リレー解説に寄せて：強化学習研究の歩み」，*計測と制御*，Vol.52, No.1, pp.68-71, 2013.

[20] 木村元：「強化学習最近の発展―第 1 回：強化学習の基礎」，*計測と制御*, Vol.52, No.1, pp.72-77, 2013.

[21] Grefenstette J.J.: “Credit assignment in rule discovery systems based on genetic algorithms”, *Machine Learning*, Vol. 3, pp. 225-245, 1988.

[22] 宮崎和光，木村元，小林重信：「Profit Sharingに基づく強化学習の理論と応用」，*人工知能学会誌*, Vol.14, No.5, pp.800-807, 1999.

[23] Barto A.G., Sutton R.S., and Anderson C.W.: “Neuron-like adaptive elements that can solve difficult learning control problems”, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol.13, No.5, pp.834-846, 1983.

[24] Sutton R.S.: “Learning to predict by the method of temporal difference”, *Machine Learning*, Vol. 3, pp. 9-44, 1988.

[25] Konda V.R., and Tsitsiklis J.N.: “Actor-critic algorithms”, *Advances in Neural Information Processing*, Vol. 12, pp. 1008-1014, 2000.

[26] Dayan P.: “The convergence of TD (λ) for general λ”, *Machine Learning*, Vol. 8, No. 3-4, pp. 341-362, 1992.

[27] Dayan P., and Sejnowski T.J.: “TD (λ) convergences with probability 1”, *Machine Learning*, Vol. 14, No. 3, pp. 295-301, 1994.

[28] Watkins C.J.C.H, and Dayan P.: “Technical note: Q-learning”, *Machine Learning*, Vol. 8, No. 3-4, pp. 279-292, 1992.

[29] Rummery G.A., and Niranjan M.: “On-line Q-learning using connectionist systems”, Technique Report, CUED/F-INFENG/TR166, Cambridge University, 1994.

[30] 足立修一：「システム同定・推定理論のダイナミクス」，*計測と制御*, Vol.52, No.4, pp.368-373.

[31] Sutton R.S.: “Generalization in reinforcement learning: Successful examples using sparse coarse coding”, *Advances in Neural Information Processing Systems*, Vol. 8, pp. 1038-1044, 1996.

[33] Sutton R.S.: “Integrated architectures for learning, planning, and reacting based on approximating dynamic programming”, *Proceedings of the 7th International Conference on Machine Learning*, pp. 216-224, 1990.

[34] Singh S.P., Jaakkola T., Jordan M.I.: “Reinforcement learning with soft state aggregation”, *Advances in Neural Information Processing Systems*, Vol.7, pp. 361-368, 1995.

[35] Sutton R.S., McAllester D., Singh S., and Mansour Y.: “Policy gradient methods for reinforcement learning with function approximation”, *Advances in Neural Information Processing Systems*, Vol. 12, pp. 1057-1063, 2000.

[36] 木村元：「ランダムタイリングとGibbs-samplingを用いた多次元状態―行動空間における強化学習」，*計測自動制御学会論文集*, Vol.42, No.12, pp.1336-1343, 2006.

[37] Kobayashi K., Mizuno S., Kuremoto T., and Obayashi M.: “A reinforcement learning system based on state space construction using fuzzy ART”, *Proceedings of SICE Annual Conference*, pp.3653-3658, 2005.

[38] Kuremoto T., Obayashi M., Yamamoto A., and Kobayashi K.: “Neural Prediction of Chaotic Time Series Using Stochastic Gradient Ascent Algorithm”, *Proceedings of the 35th ISCIE International Symposium on Stochastic Systems Theory and Its Applications (SSS’03)*, pp.17-22, 2003.

[39] Kuremoto T., Obayashi M., Yamamoto Y., and Kobayashi, K.: “Predicting chaotic time series by reinforcement learning”, in *Proceedings of 2nd International Conference on Computational Intelligence, Robotics, and Autonomous Systems (CIRAS’03),* 2003.

[40] Kuremoto T., Obayashi M., and Kobayashi K.: “Nonlinear prediction by reinforcement learning”, *Lecture Note in Computer Science,* Vol. 3644, pp.1085-1094, 2005.

[41] Kuremoto, T., Obayashi, M., and Kobayashi, K.: “Forecasting Time Series by SOFNN with Reinforcement Learning”, *The 27th Annual International Symposium on Forecasting (ISF2007)*, Neural Forecasting Competition (NN3), 2007.

[42] Kuremoto, T., Obayashi, M., and Kobayashi, K.: “Neural Forecasting Systems”. In *Reinforcement Learning, Theory and Applications (ed. Cornelius Weber, Mark Elshaw and Norbert Michael Mayer)*, *Advanced Robotic Systems*, Chapter 1, pp. 1-20, InTech, 2008.

[43] Obayashi M., Kuremoto T., Kobayashi K.: “A self-organized fuzzy-neuro reinforcement leaning system for continuous state space for autonomous robots”, *Proceedings of International Conference on Computational Intelligence for Modeling, Control, and Automation (CIMCA’08)*, pp. 552-559, 2008.

[44] Obayashi M., Iseki A., and Umesako K.: “Self-organized reinforcement learning using fuzzy inference for stochastic gradient ascent method”, *Proceedings of the International Conference on Control, Automation and Systems (ICCAS2011)*, pp.735-738, 2001.

[45] Umesako K., Obayashi M., and Kobayashi K.: “Mobile robot control using self-organized fuzzy reinforcement learning system”, *Proceedings of International Symposium on Advanced Control of Industrial Processes*, pp. 513-519, 2002.

[46] 梅迫公輔，大林正直，小林邦和：「自己組織化型ファジィ強化学習システム」，*計測自動制御学会論文集*，Vol.39, No.7, pp.699-701, 2003.

[47] Obayashi M., Narita K., Kuremoto T., and Kobayashi K.: “A reinforcement learning system with chaotic neural networks-based adaptive hierarchical memory structure for autonomous robots”, *Proceedings of International Conference on Control, Automation and Systems (ICCAS2008)*, pp. 69-74, 2008.

[48] 梅迫公輔，大林正直，小林邦和：「適応探索法を用いた強化学習」，*電気学会論文誌C*，Vol.122, No.3, pp.374-380, 2002.

[49] Obayashi M., Umesako K., Oda T., Kobayashi K., Kuremoto T.: “Actor-Critic reinforcement learning system with time-varying parameters”, *Proceedings of the International Conference on Control, Automation and Systems (ICCAS2003)*, pp.138-141, 2003.

[50] Umesako K., Obayashi M., and Kobayashi K.: “Evolutionary and time-varying reinforcement learning system based on overlap of rules”, *Proceedings of 6th Japan-France Congress on Mechatronics and 4th Asia-Europe Congress on Mechatronics*, pp. 202-207, 2003.

[51] 梅迫公輔，大林正直，小林邦和：「時変パラメータを持つ進化的強化学習システム」，*電気学会論文誌C*，Vol.124, No.7, pp.1478-1483, 2002.

[52] 小川長久，大林正直，小林邦和，呉本尭：「免疫回路網式強化学習」，*計測自動制御学会論文集*，Vol.43, No.6, pp.525-527, 2007.

[53] Kogawa N., Obayashi M., Kobayashi K., Kuremoto T.: “A reinforcement learning method based on immune network adapted to semi-Markov decision process”, *Proceedings of 13th International Symposium on Artificial Life and Robotics (AROB2008)*, pp. 63-66, 2008.

[54] Kogawa N., Obayashi M., Kobayashi K., Kuremoto T.: “A reinforcement learning method based on immune network adapted to semi-Markov decision process”, *Artificial Life and Robotics*, Vol. 13, No. 2, pp. 538-542, 2009.

[55] Lovejoy W.S.: “A survey of algorithmic methods for partially observed Markov decision process”, *Annals of Operation Research*, Vol. 28, pp. 47-66, 1991.

[56] 木村元，山村雅幸，小林重信：「部分観測マルコフ決定過程下での強化学習：確率的傾斜法による接近」，*人工知能学会誌*，Vol.11, No.5, pp.761-768, 1996.

[57] 木村元，L. P. Kaelbling：「部分観測マルコフ決定過程下での強化学習」，*人工知能学会誌*，Vol. 12, No.6, pp.822-830, 1997.

[58] 宮崎和光，荒井幸代，小林重信：「POMDPs環境下での決定的政策の学習」，*人工知能学会誌*，Vol.14, No.1, pp.148-156, 1997.

[59] 渋谷長史：「部分観測マルコフ決定過程と強化学習」，*計測と制御*，Vol.52，No.4, pp.374-380, 2013.

[60] 渡辺澄夫，萩原克幸，赤穂昭太郎，本村陽一，福水健次，岡田真人，青柳美輝：「学習システムの理論と実現」，*森北出版*，2005.

[61] Obayashi M., Yamada K., Kuremoto T., Kobayashi K.: “A robust reinforcement learning system using sliding mode control with state variable filter”, *Proceedings of 2009 CACS International Automatic Control Conference*, 2009.

[62] Obayashi M., Nakahara N., Kuremoto T., Kobayashi K.: “A robust reinforcement learning using concept of slide mode control”, *Artificial Life and Robotics*, Vol. 13, No. 2, pp. 526-530, 2009.

[63] 牧野吉宏，大林正直，呉本尭，小林邦和：「間接適応型自己構造ファジィニューラルネットワーク制御システム」，*電気学会論文誌C*，Vol.130, No.10, pp.1882-1887, 2010.

[64] 中野一宏，大林正直，呉本尭，小林邦和：「部分的未知構造を持つ非線形システムのた

めのロバスト強化学習制御計設計法」，*電気学会論文誌C*，Vol.130, No.11, pp.2090-2091, 2010.

[65] Uchiyama S, Obayashi M., Kuremoto T., Kobayashi K.: “Robust reinforcement learning system with H∞ tracking performance conpensator”, *Proceedings of the International Conference on Control, Automation and Systems (ICCAS2011)*, pp.248-253, 2011.

[66] 内山祥吾，大林正直，呉本尭，小林邦和：「H∞追従性能補償器を備えたリアルタイム強化学習制御システム」，*電気学会論文誌C*，Vol.132, No.6, pp.1008-1015, 2012.

[67] Asada M., Noda S., Tawaratumida S., and Hosoda S.: “Purposive behavior acquisition for a real robot by vision-based reinforcement learning”, *Machine Learning*, Vol. 23, pp. 279-303, 1996.

[68] Obayashi M., Oda T., Kobayashi K., Kuremoto T., and Kitano H.: “Reinforcement Learning System with Time Varying Parameters Using Neural Network”, *Proceedings of the 35th ISCIE International Symposium on Stochastic Systems Theory and Its Applications (SSS’03)*, pp.11-16, 2003.

[69] Nakano K., Obayashi M., Kobayashi K., and Kuremoto T.: “Cooperative behavior acquisition for multiple autonomous mobile robots”, *Proceedings of the 10th International Symposium on Artificial Life and Robotics (AROB2005)*, pp.543-546, 2005.

[70] Kuremoto T., Hano T., Kobayashi K., Obayashi M.: “For Partner Robots: A Hand Instruction Learning System Using Transient-SOM”, *Proceedings of The 2nd International Conference on Natural Computation and The 3rd International Conference on Fuzzy Systems and Knowledge Discovery (ICNC '06-FSKD'06*), pp.403-414, 2006.

[71] Kuremoto, T., Hano, T., Kobayashi, K., and Obayashi, M.: “Robot Feeling Formation Based on Image Features”, *Proceedings of International Conference on Control, Automation and Systems (ICCAS2007)*, pp.758-761, 2007.

[72] Kobayashi K., Nakano K., Kuremoto T., and Obayashi M.: “Cooperative Behavior Acquisition of Multiple Autonomous Mobile Robots by an Objective-based Reinforcement Learning System”, *Proceedings of International Conference on Control, Automation and Systems (ICCAS2007)*, pp.777-780, 2007.

[73] 羽野ともえ，呉本尭，小林邦和，大林正直：「Transient-SOM を用いた手画像命令学習システム」，*計測自動制御学会論文集*，Vol.43, No.11, pp.1004-1006, 2007.

[74] Kollar T., and Roy N.: “Trajectory optimization using reinforcement learning for map exploration”, *The International Journal of Robotics Research*, Vol. 27, No. 2, pp.175-196, 2008.

[75] Kuremoto, T., Komoto, T., Kobayashi, K., and Obayashi, M.: “A Voice Instruction Learning System Using PL-T-SOM”, *Proceedings of the 2nd International Conference on Image and Signal Processing (CISP2009)*, pp.4294-4299, 2009.

[76] Feng L.-B., Obayashi M., Kuremoto T., Kobayashi K.: “An intelligent control system

construction using high-level time Petri net and reinforcement learning", *Proceedings of the International Conference on Control, Automation and Systems (ICCAS2010)*, pp.535-539, 2010.

[77] Feng L.-B., Obayashi M., Kuremoto T., Kobayashi K.: "Optimization and verification for a robot control system on learning Petri net model", *Lecture Note in Electronic Engineering*, Vol. 133, pp.815-823, 2011.

[78] Kuremoto T., Yamane T., Feng L.-B., Kobayashi K., and Obayashi M.: "A human-machine interaction system: A voice command learning system Using PL-G-SOM", *Proceedings of 2011 International Conference on Industrial Engineering and Management (IEEE-IEM 2011)*, pp.83-86, 2011.

[79] Kuremoto T., Otani, T., Feng L., Kobayashi K., and Obayashi M.: "A hand image instruction learning system using PL-G-SOM", *Proceedings of the 12th International Conference on Artificial Intelligence (ICAI 2012 / WORLDCOMP 2012*), 2012.

[80] Obayashi, M., Takuno T., Kuremoto T., and Kobayashi K.: "An emotional model embedded reinforcement learning system", *Proceedings of the IEEE International Conference on System, Man, and Cybernetics (IEEE SMC 2012)*, 2012.

[81] Kuremoto, T., Hashiguchi K. Morisaki K., Watanabe S., Kobayashi, K., Obayashi M.: "Multiple action sequence learning and automatic generation for a humanoid robot using RNNPB and reinforcement learning", *Journal of Software Engineering and Applications*, Vol.5, pp.128 -133, 2012.

[82] 飯間等, 黒江康明：「エージェント間の情報交換に基づく群強化学習法」, *計測自動制御学会論文集*, Vol.42, No.11, pp.1244-1251, 2006.

[83] Iima H., and Kuroe Y.: "Swarm reinforcement learning algorithm based on Sarsa method", *Proceedings of SICE Annual Conference*, pp. 2045-2049, 2008.

[84] 小林邦和，中野浩二，呉本尭，大林正直:「状態予測型強化学習システム」，*電気学会論文誌C*，Vol.128, No.8, pp.1303-1311, 2008.

[85] Kuremoto T, Obayashi M., Kobayashi K., Adachi H., and Yoneda K.: "A reinforcement learning system for swarm behaviors", *Proceedings of IEEE World Congress on Computational Intelligence / International Joint Conference on Neural Networks (WCCI/IJCNN)*, pp.3710-3715, 2008.

[86] Kuremoto T, Obayashi M., Kobayashi K., Adachi H., and Yoneda K.: "A neuro-fuzzy learning system for adaptive swarm behaviors dealing with continuous state space", *Lecture Notes in Artificial Intelligence*, Vol. 5227, pp.675-683, 2008.

[87] Kuremoto T, Obayashi M., and Kobayashi K.: "Adaptive swarm behavior acquisition by a neuro-fuzzy system and reinforcement learning algorithm", *International Journal of Intelligent Computing and Cybernetics*, Vol. 2, No. 4, pp. 724-744, 2009.

[88] Kuremoto T., Yamano Y., Obayashi M., and Kobayashi K.: “An improved internal model for swarm formation and adaptive swarm behavior acquisition”, *Journal of Circuits, Systems, and Computers*, Vol. 18, No. 8, pp. 1517-1531, 2009.

[89] Kobayashi K., Obayashi M., Kuremoto T.: “Objective-based Reinforcement Learning System for Cooperative Behavior Acquisition”, In *Machine Learning*, *Advanced Robotic Systems*, Chapter 14, pp.233-244, InTech, 2010.

[90] Kobayashi, K., Kanehira, R., Kuremoto, T., and Obayashi, M.: “An Action Selection Method Based on Estimation of Other's Intention in Time-Varying Multi-Agent Environments”, *Lecture Note in Computer Science (LNCS)*, Vol. 7064, pp.76-85, 2011.

[91] Kobayashi K., Kurano T., Kuremoto T., Obayashi M.: “Cooperative behavior acquisition in multi-agent reinforcement learning system”, *Lecture Notes in Computer Science*, Vol. 7665, pp.537-544, 2012.

[92] Kuremoto T., Yamano Y. Feng L.-B, Kobayashi K., and Obayashi M.: “A fuzzy neural network with reinforcement learning algorithm for swarm learning”, *Lecture Notes in Electronic Engineering (LNEE)*, Vol.144, pp.101-108, 2011.

[93] 呉本尭，山野祐樹，馮良炳，小林邦和，大林正直：「ニューロファジィ型強化学習システムを用いた群行動の獲得」，*電気学会論文誌C*，Vol.133, No.5, pp.1076-1085, 2013.

[94] Kuremoto T., Obayashi M., and Kobayashi K.: “Neuro-fuzzy systems for autonomous mobile robots”, *Horizons in Computer Science Research*, Vol. 8, pp.67-90, 2013

[95] Kuremoto T., Kobayashi K., and Obayashi M.: “Nonlinear prediction by reinforcement learning”, *Lecture Notes in Computer Science*, Vol.3644, pp.1085-1094, 2005.

[96] Feng L.-B., Obayashi M., Kuremoto T., and Kobayashi K.: “A learning Petri net model”, *IEEJ Transactions on EEE,* Vol.7, No.3, pp.274-282, 2012.

[97] Feng L.-B., Obayashi M., Kuremoto T., and Kobayashi K.: “QoS optimization for Web services Composition based on reinforcement learning”, *International Journal of Innovative Computing, Information and Control*, Vol.9, No.6, pp.1-10, 2013.

[98] Zadeh L.A.: “Fuzzy sets”, *Information and Control*, Vol.8, pp.383-353, 1965.

[99] 谷萩隆嗣, 萩原将文, 山口亨：「ニューラルネットワークとファジィ信号処理」，*コロナ社*, 1998.

[100] 田中雅博：「ソフトコンピューティング入門」，*科学技術出版*, 1999.

[101] Takagi T., and Sugeno M.: “Fuzzy identification of systems and its applications to modeling and control”, *IEEE Transactions on Systems, Man and Cybernetics*, Vol.15, No.1, pp.116–132, 1985.

[102] M. Sugeno and Kang G.T.: “Structure identification of fuzzy model”, *Fuzzy Sets and Systems*, Vol.28, pp.15-33, 1988.

[103] Yen J., Wang W., and Gillespie C.W.: “Improving the interpretability of TSK fuzzy models

by combining global learning and local learning", *IEEE Transactions on Fuzzy Systems*, Vol. 6, No.4, pp.530-537, 1998.

[104] Jang J-S.R.: "ANFIS: Adaptie-network-based fuzzy inference system", *IEEE Transactions on Systems, Man and Cybernetics,* Vol. 23, No.3, pp.665-685, 1993.

[105] Jang J-S.R., and Sun C.T.: "Functional equivalence between radial basis function networks and fuzzy inference systems", *IEEE Transactions on Neural Networks*, Vol. 4, No.1, pp.156–159, 1993.

[106] Lin C.T., Lin C.J., Lee C.S.G.: "Fuzzy adaptive learning control network with on-line neural learning", *Fuzzy Sets and Systems*, Vol.71, No.1, pp.25–45, 1995.

[107] Jouffe L.: "Fuzzy inference system learning by reinforcement learning", *IEEE Transactions on System, Man and Cybernetics-B*, Vol. 28, No. 3, pp.338-355, 1998.

[108] Wang X.S., Cheng Y.H., and Yi J.Q.: "A fuzzy actor–critic reinforcement learning network", *Information Sciences*, Vol. 177, pp.3764-3781, 2007.

[109] Derhami V., Majd V.J., and Ahmadabadi M.N.: "Exploration and exploitation balance management in fuzzy reinforcement learning", *Fuzzy Sets and Systems*, Vol. 161, No. 4, pp. 578－595, 2010.

[110] Reynolds C.W.: "Boids background and update," http://www.red3d.com/cwr/boids/.

[111] Reynolds C.W.: "Flocks, herds, and schools: A distributed behavioral model, in computer graphics", *SIGGRAPH'87 Conference Proceedings*, Vol. 21, No. 4, pp. 25-34, 1987.

[112] Kennedy J., and Eberhart R.C.: "Particle swarm optimization", *Proceedings of the IEEE International Conference on Neural Networks*, pp.1942-1948, 1995.

[113] Kennedy J., Eberhart R.C., and Shi Y.: *Swarm Intelligence*, San Francisco, Morgan Kaufmann Publishers, 2001.

[114] Dorigo M., Maniezzo V., Colorni A.: "Ant system: optimization by a colony of cooperating agents", *IEEE Transactions on Systems, Man, and Cybernetics-B*, Vol.26, No.1, pp.29-41, 1996.

[115] Dorigo M., and Caro G.D.: "Ant colony optimization: a new meta-heuristic", *Proceedings of Congress on Evolutionary Computation*, pp. 1470-1477, 1999.

[116] A. Jababaie, J. Lin, and A. S. Morse: "Cooperation of groups of mobile autonomous agents using nearest neighbor rules", *IEEE Transaction on Automatic Control*, Vol. 48, No. 6, pp. 988-1001, 2003.

[117] Moreau L.: "Stability of multiagent systems with time-dependent communication links", *IEEE Transaction on Automatic Control,* Vol. 50, No. 2, pp.160-182, 2005.

[118] Hou Z.G., Cheng L., Tan M.: "Decentralized robust adaptive control of the multi-agent system consensus problem using neural networks", *IEEE Transactions on Systems, Man and Cybernetics-B*, Vol. 39, No. 3, pp. 636-647, 2009.

[119] 牧野貴樹：「強化学習最近の発展―第 2 回：探索と利用のトレードオフとベイズ環境モデル」，*計測と制御*，Vol.52, No.2, pp.154-161, 2013.

[120] Doya K.: “Metalearning and neuromodulation”, *Neural Networks*, Vo1.15, Issue 4-6, pp.495-506, 2002.

[121] Ishii S., Yoshida W., Yoshimoto J.: “Control of exploitation-exploration meta-parameter in reinforcement learning”, *Neural Networks*, Vol.15, Issue 4-6, pp.665-687, 2002.

[122] Ishida F., Sasaki T., Sakaguchi Y., Shimai H.: “Reinforcement-learning agents with different temperature parameters explain the variety of human action-selection behavior in a Markov decision process task”, *Neurocomputing*, Vol. 72, Issue 7-9, pp.1979-1984, 2009.

[123] Kobayashi K., Mizoue H., Kuremoto T., and Obayashi M.: “A meta-learning method based on temporal difference error”, *Lecture Notes in Computer Science*, Vol. 5863, pp. 530-537, 2009.

[124] 溝上裕之，小林邦和，呉本尭，大林正直：「TD 誤差に基づく強化学習のメタパラメータ学習法」，*電気学会論文誌 C*，Vol.129, No.9, pp.1730-1736, 2009.

[125] 堀内匡，藤野昭典，片井修：「経験強化を考慮した Q-Learning の提案とその応用」，*計測自動制御学会論文集*，Vol.35, No.5, pp.645-653, 1999.

[126] 佐藤誠，木村元，小林重信：「報酬の分散を推定する TD アルゴリズムと Mean-Variance 強化学習の提案」，*人工知能学会論文誌*，Vol.16, No.3, pp.353-362, 2001.

[127] 植野剛，前田新一，川鍋一晃：「強化学習最近の発展―第 3 回：統計学習の観点から見た TD 学習」，*計測と制御*，Vol.52, No.3, pp.277-283, 2013.

[128] Kuremoto T., Obayashi M., Kobayashi K., and Feng L.-B.: “Autonomic behaviors of swarm robots driven by emotion and curiosity”, *Lecture Notes in Bioinformatics*, Vol. 6330, pp. 541-547, 2010.

[129] Kuremoto T., Obayashi M., Kobayashi K., and Feng L.-B.: “An improved internal model of autonomous robots by a psychological approach”, *Cognitive Computation*, Vol. 3, No. 4, pp. 501-509, 2011.

[130] Watada S., Obayashi M., Kuremoto T., and Kobayashi K.: “A new decision-making system of an emotional agent based on emotional models in multi-agent system”, *Proceedings of the 18th International Symposium on Artificial Life and Robotics (AROB2013)*, pp. 452-455, 2013.

[131] Kuremoto T., Tsurusaki T., Kobayashi K., Mabu S., and Obayashi M.: “A model of emotional intelligent agent for cooperative goal exploration”, *Lecture Notes in Computer Science*, Vol. 7995, pp.21-30, 2013.

[132] Kuremoto T., Tsurusaki T., Kobayashi K., Mabu S., and Obayashi M.: “An improved reinforcement learning system using affective factors”, *Robotics*, Vol. 2, No. 3, pp.149-164, (doi:10.3390/robotics2030149) 2013

付録 A

Sarsa 学習アルゴリズム[15][29][31]

Step 1 任意値で状態―行動価値関数 $Q(s_t,a_t)$ を初期化する。$t=0$、学習率 $0<\alpha_t\le 1$ と減衰率 $0\le\gamma\le 1$ を設定する。

Step 2 探索試行（cycle）ごとに以下を繰り返す。

状態 s_t を初期に戻す。

時刻 t ごとに以下を s_t が最終状態になるまで繰り返す。

$a_t \leftarrow s_t$ （政策 $\pi(s_t)$ を用いて得られる行動を a_t へセット）

行動 a_t を実行し、報酬 r_t と次状態 s_{t+1} を観測する。

$a_{t+1} \leftarrow s_{t+1}$ （政策 $\pi(s_{t+1})$ を用いて得られる行動を a_{t+1} へセット）

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left[r_t + \gamma\cdot Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)\right] \qquad (A.1)$$

$s_t \leftarrow s_{t+1}$ （次状態へ）

付録 B

Q 学習アルゴリズム[15][28][31]

Step 1 任意値で状態―行動価値関数 $Q(s_t,a_t)$ を初期化する。 $t=0$ 、学習率 $0<\alpha_t\le 1$ と減衰率 $0\le\gamma\le 1$ を設定する。

Step 2 探索試行（cycle）ごとに以下を繰り返す。

状態 s_t を初期に戻す。

時刻 t ごとに以下を s_t が最終状態になるまで繰り返す。

$a_t \leftarrow s_t$ （政策 $\pi(s_t)$ を用いて得られる行動を a_t へセット）

行動 a_t を実行し、報酬 r_t と次状態 s_{t+1} を観測する。

$\max_{a'\in A} Q(s_{t+1},a')$ （次状態の最大の $Q(s_{t+1},a_{t+1})$ を観測する）

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha\left[r_t+\gamma\cdot\max_{a'\in A}Q(s_{t+1},a')-Q(s_t,a_t)\right] \tag{B.1}$$

$s_t \leftarrow s_{t+1}$ （次状態へ）

付録 C

SGA 学習[44]を用いた自己組織化型ファジィ強化学習システム[46]

自己組織化ファジィネット(Fuzzy net)を観測状態の認知モジュールとし、その出力を重み付けて、行動選択の確率方策関数のパラメータとして用いる強化学習システム[46]の構成を図 C.1 に示す

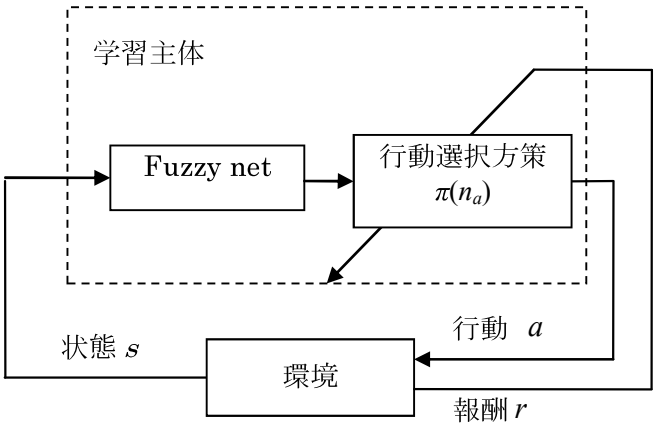


図 C.1 自己組織化型ファジィ強化学習システム[46]の構成

行動選択のための確率方策関数 $\pi(n_a)$ が(C.1)式の Adaptive PDF [48]を用いる場合、ファジィネットの出力と結合したのは方策関数の内部変数（パラメータ）p, μ, β となり、それぞれ、(C.2)式、(C.3)式、(C.4)式で与える。

$$\pi(n_a)=\begin{cases} p\beta e^{\beta(x-\mu)} & (n_a \le \mu) \\ (1-p)\beta e^{-\beta(x-\mu)} & (n_a > \mu) \end{cases}, \tag{C.1}$$

$$p=\frac{\sum_k w_p^k \phi_t^k(\mathbf{x}(t))}{\sum_k \phi_t^k(\mathbf{x}(t))}, \tag{C.2}$$

$$\mu=\frac{\sum_k w_\mu^k \phi_t^k(\mathbf{x}(t))}{\sum_k \phi_t^k(\mathbf{x}(t))}, \tag{C.3}$$

$$\mu = \frac{\sum_k w_\mu^k \phi_t^k(\mathbf{x}(t))}{\sum_k \phi_t^k(\mathbf{x}(t))}, \tag{C.3}$$

但し、ここで、n_a は $\pi(n_a)$ に従う乱数で、$\phi_t^k(\mathbf{x}(t))$ は(2.7)式のファジィネットの出力で、$w_p^k, w_\mu^k, w_\beta^k$ はファジィネットと方策関数パラメータの結合荷重である。$k = 1,2,..K_t$ はファジィルール番号を表す。

すべての状態に対して最大の報酬を獲得するよう適応行動を選択するため、Stochastic Gradient Ascent (SGA)学習[44]を用いて、方策関数の内部変数（パラメータ）p, μ, β 及びファジィネットと方策関数の結合荷重 $w_p^k, w_\mu^k, w_\beta^k$ を修正する。SGA 学習アルゴリズムを図 C.2 に示す。

Step 1　環境の観測 x_t を受け取る。
Step 2　方策関数 $\pi(a_t, w, x_t)$ の確率で行動を実行する。ここで、w は内部変数（パラメータ）である。
Step 3　環境から報酬 r_t を受け取る。
Step 4　内部変数 w のすべての要素 ϖ_i について以下の Characteristic Eligibility $e_i(t)$ と $\bar{\bar{D}}_i(t)$ を求める。但し γ は割引率である $(0 \le \gamma < 1)$。

$$e_i(t) = \frac{\partial}{\partial \varpi_i} \ln\{\pi(a_t, w, x_t\}$$

$$\bar{\bar{D}}_i(t) = e_i(t) + \gamma \bar{\bar{D}}_i(t-1)$$

Step 5　以下の式を用いてを求める。

$$\Delta \varpi_i(t) = (r_t - b)\bar{\bar{D}}_i(t)$$

但し、b は定数である。
Step 6　方策の改善：以下の式で w を更新

$$\Delta w(t) = (\Delta \varpi_1(t), \Delta \varpi_2(t), .., \Delta \varpi_i(t)..),$$
$$w \leftarrow w + \alpha \Delta w(t)$$

但し、α は非負の学習定数である。
Step 7　時間ステップをへ進めて、Step 1 へ戻る。

図 C.2 SGA 学習アルゴリズム[44]

なお、4.1.4 節の目標探索シミュレーションにおいて、個体の取り得る行動 Action 1～Action 4 （上下左右方向へそれぞれ 1 マス移動）は図 C.3 に示す確率 $\pi(n_a)$ で決定される。図 C.3 において、パラメータ p=0.25, μ=0.0, β=0.5 であり、θ_1, θ_2 は閾値である。

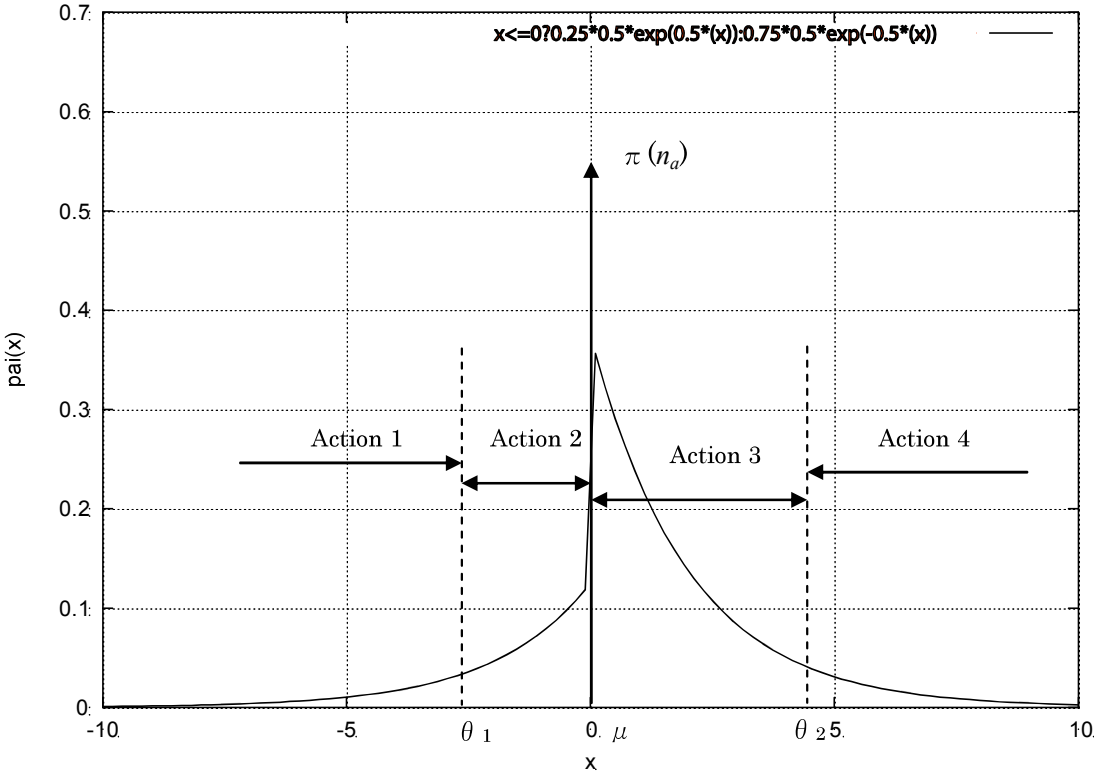


図 C.3 Adaptive PDF（適応確率密度分布関数）[48]

SGA 学習を用いた自己組織化ファジィ強化学習システムの方策改善アルゴリズムを図 C.4 に示す。

Step 1　環境の観測 x_t を受け取る。

Step 2　2 章で述べた SOFNN を構築する。ファジィネットと方策関数との結合荷重 $w_p^k, w_\mu^k, w_\beta^k$ を初期化する。

Step 3　(C.1)式～(C.4)式で方策関数を構成する。

Step 4　以下の式で n_a 計算し、その値に対応した行動を選択する（図 C.3 参照）。

$$n_a = \begin{cases} \dfrac{1}{\beta}\ln(\dfrac{z}{p}) + \mu & (0 \le z \le p) \\ -\dfrac{1}{\beta}\ln(\dfrac{1-z}{1-p}) + \mu & (p < z \le 1.0) \end{cases}$$

但し、z は[0,1]の一様乱数である。

Step 5　環境より報酬を受け取る。

Step 6　SGA により学習パラメータを更新する。

Step 7　Step 1 へ戻る。

図 C.4 SGA 学習を用いた自己組織化型ファジィ強化学習システムの学習アルゴリズム

（終わり）