平成　19 年度

卒 業 研 究 論 文

題　　目　　群の中の個体の知能の発達

提 出 者　　室蘭工業大学 情報工学科

氏　名　　木島　康隆

学籍番号　　1623025

提出年月日　　平成　20 年　2 月　13 日

室蘭工業大学
情報工学科

目次

第１章　序論 1
1.1　本研究の背景 1
1.2　　従来研究 2
1.2.1　ロボット単体の知能の発達に関する研究 2
1.2.2　群ロボットを扱った研究 5
1.3 本研究の目的 6
1.4　本論文の構成 7

第２章　コミュニケーションによる知能の発達 9
2.1　コミュニケーションによる個体知能発達システム 9
2.1.1　システムの概要 9
2.1.2　他者から得られる情報の考察 10
2.2　コミュニケーションについての考察 12
2.2.1　コミュニケーションの条件 12
2.2.2　コミュニケーションする情報 13
2.3　学習法をコミュニケーション情報とした個体知能の発達 13

第３章　強化学習 16
3.1　強化学習の概要 16
3.1.1　強化学習の特徴 16
3.1.2　環境とエージェントとの相互作用とエージェントの目的 16
3.1.3　強化学習の構成要素 17
3.1.4　　強化学習の流れ 18
3.2　行動選択手法 18
3.2.1　greedy 法 19
3.2.2　ε - greedy 法 19
3.2.3　softmax 法 19
3.2.4　追跡手法 20
3.2.5　強化比較法 21
3.3　行動評価手法 22
3.3.1　標本平均手法 22
3.3.2　加重平均手法 23
3.3.2　Q 学習法 23
3.4　まとめ 24

第４章　強化学習を適用したコミュニケーションによる学習システム 25
4.1　学習法をコミュニケーションする個体知能の発達概念 25
4.2　強化学習を適用したコミュニケーション学習システム 26
4.3　まとめ 28

第５章　Ｎ本腕バンディット問題を対象とした提案システムの有効性の検証 29
5.1　Ｎ本腕バンディット問題とは 29
5.2　実験概要 29
5.3　実験設定 32
5.4　実験結果 37
5.4.1　学習法選択の推移について 37
5.4.2　獲得報酬量について 37
5.5　実験考察 50
5.5.1　学習法選択について 50
5.5.2　平均獲得報酬量について 52
5.5　まとめ 52

第６章　結論 53
6.1　まとめ 53
6.2　今後の課題 53

謝辞 56
参考文献 57

第１章　序論

1.1　本研究の背景

実用初期のロボットは，工場のラインで動作するものが主なものであった．時代が進むにつれて，ハードウェア技術が進歩してきた．その結果，工場で動作するロボット以外にも様々な用途，形態のロボットが作られ，使われ始めるようになってきた．例えば，災害救助用ロボットといった極限環境作業用ロボット（図 1.1（a)），ペットロボットなどのホビー用ロボット（図 1.1(b)），掃除ロボット（図 1.1(c)）などの便利ロボットといったように多種多様なロボットが作られ，オフィスや家庭環境，自然環境など多種多様な環境で使われ始めている．

(a) 災害救助ロボット　援竜　　(b)ペット型ロボット　ネコロ　　(c) 清掃ロボット　RFS1

図 1.1　様々なロボット

ロボットが多種多様な環境で使われるということは，ロボットの直面する環境が多様・複雑化するということである．例えば，家庭の環境ひとつとっても，常に変化している．周辺の家具の位置や，床に散らばっている本などの物体はその時々で場所を変える．また子供やペットといった常に動き回っているものもある．このような環境は，動的かつ複雑で予想することができない．

以前ロボットが運用されていた環境は，工場の一画や研究室といった単純で変化の少ない環境が主であった．現在はオフィスや家庭環境といった複雑で変化の多い環境下で運用することが望まれるようになった．以前の単純で変化の少ない環境のときは，ロボットの直面する環境を予測することが可能であったため，ロボットの行動はすべて人間のプログラムによって生成されていた．しかし，現在の複雑で変化の多い環境では，ロボットが直面しうる全ての環境を予測し，それに対して１つ１つ適切な行動をプログラムするのは不可能である（図 1.2）．

このようなことから，現在ロボットに望まれている能力として，多様・複雑な環境下で

機能停止に追い込まれることなくタスクを遂行できる能力がある．そこで，ロボットが直面する環境に合った行動を人間がプログラムするのではなく，ロボット自身が周囲の環境情報をもとに，自ら環境に合った動きを判断し，実行する必要がある．

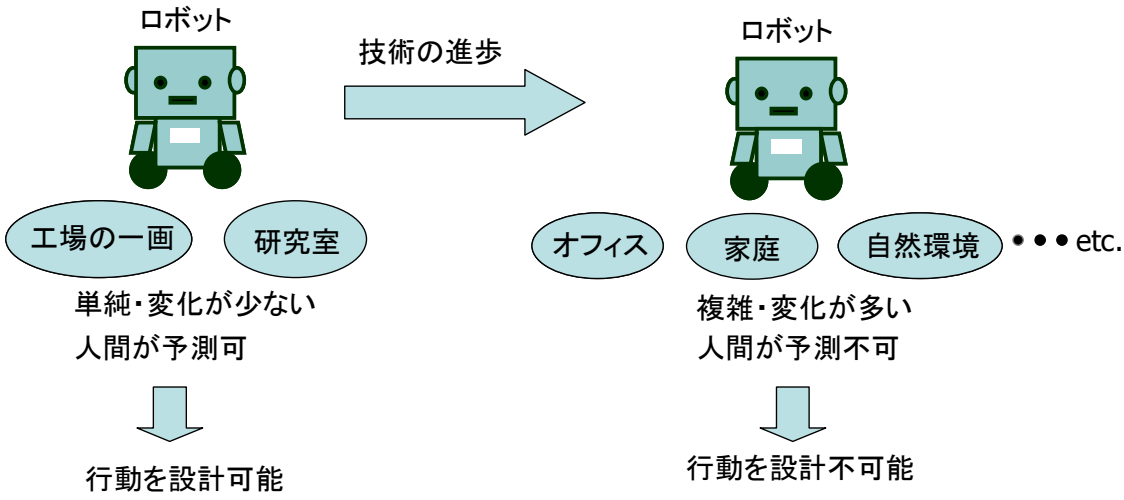


図 1.2　ロボットが直面する環境の変化と行動の設計の可否

1.2　従来研究

ロボットの環境適応の従来研究としては，以下のようなものがある．

■ 個体の知能の発展に関する研究

ロボットに人間と同じ様な学習機構を実装することによって，ロボット自身が直面する環境に合った動きを学習させる研究

■ 群ロボットを扱った研究

・協調行動をさせることで柔軟に環境の変化に適応した行動を群で行う研究

・競合学習のような群を利用して自己の知能を発達させる研究

まず，1.2.1 でロボット単体の知能について「教師あり学習」と「教師なし学習」に大別して述べる．次に 1.2.2 で群ロボットを扱った研究について述べる．

1.2.1　ロボット単体の知能の発達に関する研究

人間では，過去に行った問題とよく似たような問題に直面した時，以前に解いた経験を活かしてうまく解決することが可能である．こうした人間のような学習能力をロボットに持たせることで自身が直面する環境に適応するように学習させる，機械学習の研究が行わ

れてきた．一般に機械学習は「教師あり学習」と「教師なし学習」の２つに大別される．

教師あり学習

教師あり学習は，学習のための正解情報を与える教師が存在する学習方式である．教師あり学習の代表的な例として，ニューラルネットワークがある．

・ニューラルネットワーク

ニューラルネットワーク[1]とは，生物の神経細胞（ニューロン）をモデル化した計算素子を相互に接続したネットワーク構造のことを指す（図 1.3）．

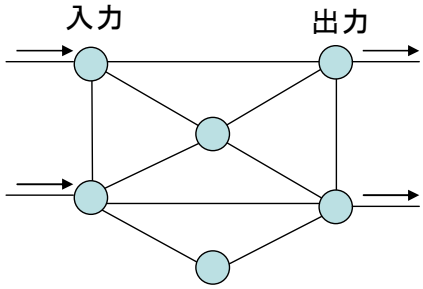


図 1.3 ニューラルネットワークの例

ニューラルネットを構成する神経細胞（ニューロン）について解説する．ニューロンは入力信号の合計値を計算し，適当な関数を通したうえで，その値が閾値を超えたら出力するという計算素子である（図 1.4）．一般に神経細胞は複数の入力を持ち，出力は１つである．神経細胞が出力する状態を「発火した」という．ニューラルネットワークの神経細胞は閾値素子として動作する．ニューロンに入る信号の値に結合加重と呼ばれる値を掛け算し，足し合わせた値が敷地を超えているかどうかで出力が変化する．

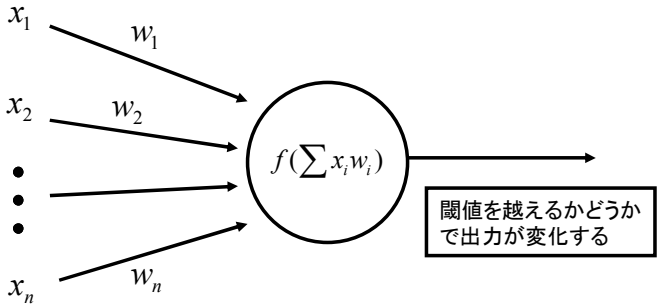


図 1.4 ニューラルネットワークにおける神経細胞（ニューロン）のモデル

ニューラルネットワークは,最終的な出力を教師データの値に近づけるようにニューラルネットワークを構成する各ニューロンの結合加重の値を増減することで学習を行う．代表的な学習アルゴリズムとしては，バックプロパゲーション法がある．

教師あり学習は，正解情報を与えられて始めて機能する．正解情報を与えるのは人間である．また，複雑で変化に富んだ環境下では，人間は何が正解かという予想がつかないため，正解情報をロボットに与えることは難しい．よって複雑で変化に富んだ環境のもとでは教師あり学習を適用するのは困難である．

教師なし学習

教師なし学習は，学習のための正解データを示す教師は存在しない．予め定められたアルゴリズムに従って学習を進めたり，環境との相互作用によって学習者（エージェント）自身が試行錯誤を通して自身にとって最適な解を学習する．教師なし学習の代表的な例として強化学習が挙げられる．

・強化学習

強化学習[2]は，環境との相互作用を通じて行動を学習する機械学習である．強化学習では，環境により報酬が与えられる．報酬は，エージェントが直面している環境下で取った行動に関して，その行動の良し悪しを数値として表したものである．その行動がよければ高い報酬を，悪ければ低い報酬または負の報酬がエージェントに与えられる．エージェントはより高い報酬を得られるようにそれぞれの状況でどのような行動をとるのが良いかを学習する．強化学習におけるエージェントは，状況認識・意思決定・行動というサイクルを繰り返す（図 1.5）．

強化学習における報酬は，正解のモデルを示すものではないため報酬は教師信号ではない．しかし，完全に教師なし学習であるとはいえない．それは，一般に報酬の算出の仕方は，人間によって設計されるものであり，環境からエージェント自身が報酬を決定するという訳ではないためである．

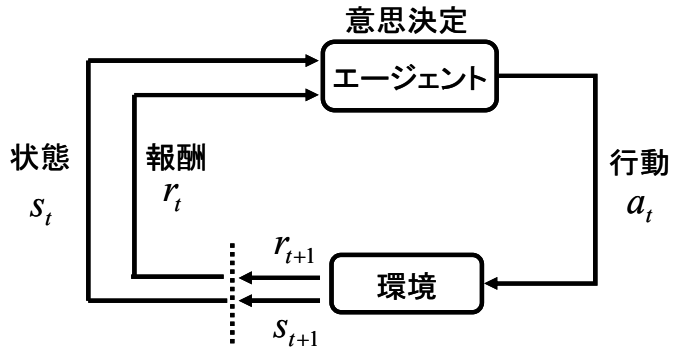


図 1.5　強化学習におけるエージェントの環境との相互作用

1.2.2　群ロボットを扱った研究

群ロボットを扱う研究では，複数台ロボットで協調して１つのタスクの処理にあたるといった協調に関する研究や，他者と競い合うことで個々の知能を発達させる競合学習といった群を利用した自己の知能の発達に関する研究が行われている．

群ロボットの協調行動に関する研究

協調行動は，郡中のロボットは単純な動きをしながら，他のロボットと協調する．群の視点で見たとき，群の動きが環境に適応したものとなる．協調行動に関する研究は，協調行動を行うために，個々のロボットを制御する手法について注目する．現在，個々のロボットを制御する手法として研究されているものは，主なものとして，集中管理型，自律分散型がある．

集中管理型では，一台のコントローラに各ロボットが獲得したセンサー情報が送られる．コントローラは送られてきた情報から各ロボットへの行動を決定し，指示する（図 1.6）．集中管理型の群ロボットの特徴として，タスクに対し統計的な方針を立てることができ，作業効率が良く完成したタスクの質が高い．欠点として，環境の変化に柔軟に対応することができないこと，群の一部が故障すると機能が著しく低下または，正常な動作ができなくなることが挙げられる[3][4].

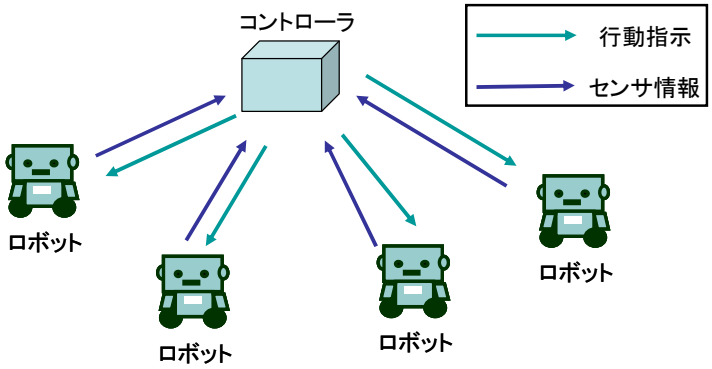


図 1.6　集中管理型

自律分散型では，集中管理型のような全てのロボットを管理するコントローラは存在せず，個々のロボットが自身の獲得するセンサー情報や他のロボットとの通信により行動を決定し，各ロボットが協調してタスクにあたる（図 1.7）．自律分散型の特徴は，個々のロボットの判断で行動するため，環境の変化に柔軟に対応することが可能であり，一部のロボットが故障しても他のロボットがある程度補うことが可能なことが挙げられる．しかし，集中管理型よりシステム全体の効率はよくない[5]．より激しい環境への適応を考えると集中管理型よりも自律分散型の方が好ましいため，現在盛んに研究がされている[3][5][6].

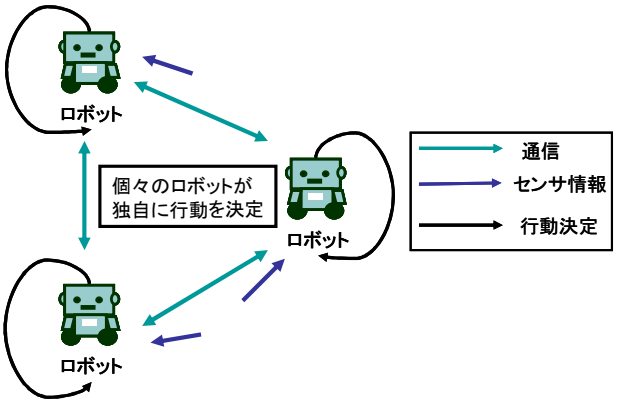


図 1.7　自律分散型

群を利用した個々のロボットの知能の発達に関する研究

前述した通り、群を利用した個々の知能の発達に関する研究には，競合学習による知能の発達がある．競合学習は，他者と競わせる中で互いの知能の発達を促すというものである．競合学習で使われるタスクの 1 つには捕食者と非捕食者の関係がある．この関係を用い，捕食者は非捕食者をうまく捕まえるような行動を学習し，非捕食者は捕食者からうまく逃げるように行動を学習する[7]．それにより，個々がそれぞれの役割を達成するための最適な行動を学習する．

1.3 本研究の目的

従来研究では，個々の知能の発達と，群を形成しての振る舞い，群を利用しての個の知能の発達というような研究が行われている．個体の知能の発達に群を利用するという研究について考えると，競合学習の研究では，競合の対象が１，２体といった少数のエージェントに限られてしまっている．また，競合学習では，エージェントは互いに敵対関係であるため，協力して個々の知能の発達をするというものではない．しかし，群を利用するならば，多くの他者と協力して多くの情報を取り入れるような方法が望ましいと考える．

人間の場合を考えると，人間は個々に試行錯誤で学習しつつも，群を形成し他者とのコミュニケーションという協力手段により情報を取り入れ自身の学習に反映している（図 1.8）．その結果，個体のみで学習するよりもより効率的に学習することができる．例えば，行ったことのない場所へ行くことを考えると，自分自身で試行錯誤によって道に迷いながらも目的地にいつかはたどり着くことはできる．しかし，途中で道行く人に目的地までの道を尋ねながら行く方がより速く目的地に到達できる．このように自分自身で試行錯誤によって解に辿り着くよりも，他者とのコミュニケーションによる情報を利用した方が効率の良い場合が多い．このような人間のコミュニケーションを用いた学習の仕方を反映したシステムをロボットに実装するのは個体の知能の発達に有効であると考える．

そこで本研究の目的として他者とのコミュニケーションによって個々の知能の発達を促進するシステムの構築を行う（図 1.9）.

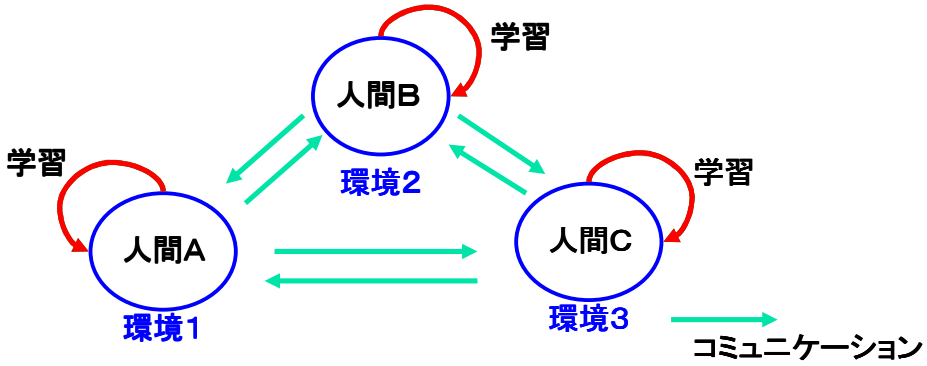


図 1.8　コミュニケーションによる学習（人間の例）

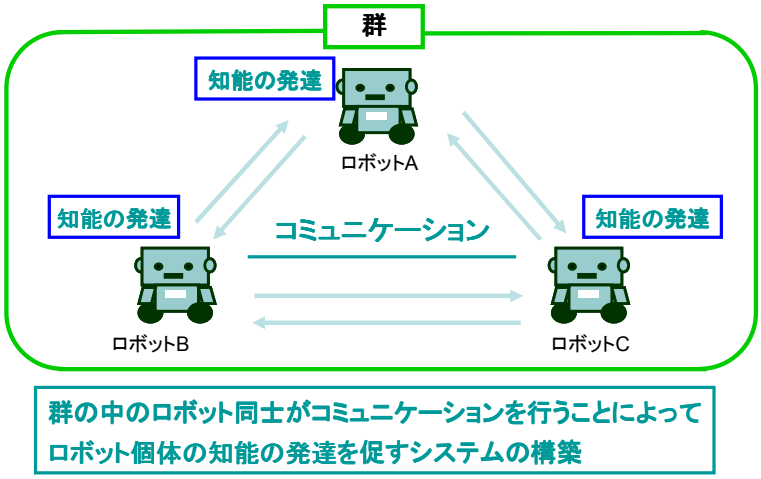


図 1.9　本研究の目的

1.4　本論文の構成

第 1 章では，本研究の背景及び従来研究について説明し，本研究の目的を示す.

第 2 章では，始めに，本研究で目指すシステムについて述べ，コミュニケーションによって得られる情報についての考察を行う．次に本論文で考えるコミュニケーションについて述べ，コミュニケーションする情報を考える．そして，その情報をコミュニケーションするシステムの概念について述べる.

第３章では，第２章で提案した手法に強化学習を適用するため，強化学習について説明する．

第４章では，第２章で提案した手法に強化学習を適用したシステムについて説明する．

第５章では，第４章で説明したシステムを用いてコミュニケーションの有効性を検証するため行った実験の内容及び結果を述べる．

第６章では本論文の結論及び今後の課題を述べる．

第２章　コミュニケーションによる知能の発達

2.1　コミュニケーションによる個体知能発達システム

2.1.1　システムの概要

本研究の目標であるコミュニケーションによる個体知能の発達を促進するシステムについて考える．知能は，学習することによって発達する．学習に必要なのは，学習対象の情報である．学習対象は自身の環境認識能力（認識できる環境状態の数）と意思決定能力（行える行動の数）からなる空間（学習空間）である．学習は学習空間内を探索し，自身の直面する環境に最適な行動を試行錯誤より見つけることである．学習空間は，ロボットの身体構造の複雑さに比例する．さらに，実環境では，エネルギーなどの問題から活動時間が限られるため，学習時間が無制限にあるわけではない．このため，限られた時間の中で広大な学習空間に対し学習を行わなくてはならない．

個体単体で学習する場合，自身の試行錯誤から得られる情報のみで学習空間から，適切な行動を探索しなくてはならない．学習空間が広い場合，自身が得る情報のみでは探索しきれない場合がある．そこで，コミュニケーションによる他者からの情報を利用することを考える．他者情報には，自身がまだ探索していない学習空間の領域に関する情報も存在する．そのような情報は自身が試行錯誤によって時間は掛かるが探索することができる．しかし，実環境では時間的制約があるため，より早く直面する環境に適した行動を獲得する方が良い．コミュニケーションによる学習は，自身が探索していない情報も他者から取り入れることができるため，より早く学習することが可能になる．つまり，コミュニケーションを行うことで，個体単体で試行錯誤により学習するよりも多くの情報を得ることが可能となる（図 2.1，2.2）．そして，より多くの情報を獲得することで，多く獲得した情報の分だけ個体単体で学習するよりも効率よく学習することが可能となる．

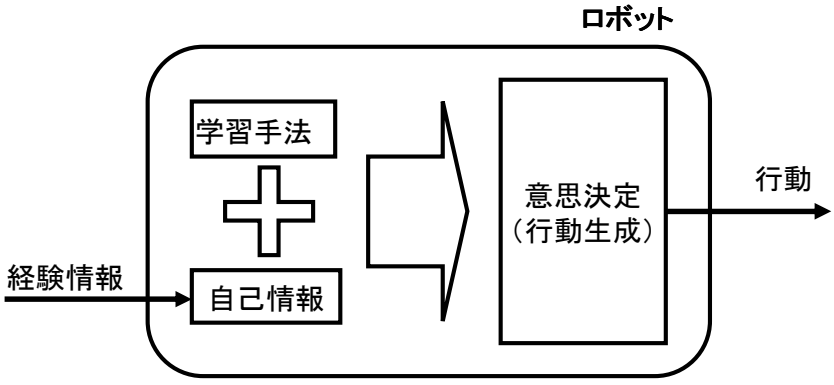


図 2.1　単体での学習

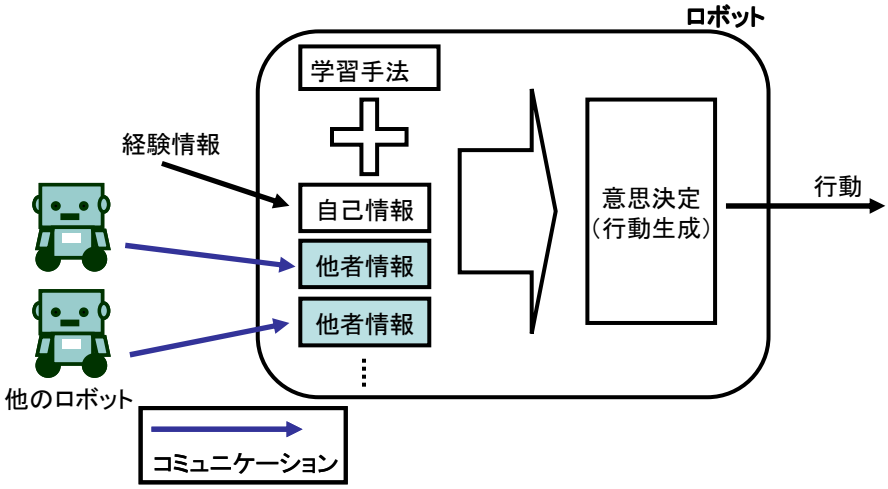


図 2.2　コミュニケーションを活用した学習

提案するシステムは，個体がより効率よく学習するために，他者とのコミュニケーションによる情報を利用するというものである．よって，個体単体でも学習を行うことが出来，他者とのコミュニケーションを用いることでより効率よく学習するようなシステムの実現を目指す（図 2.3）．

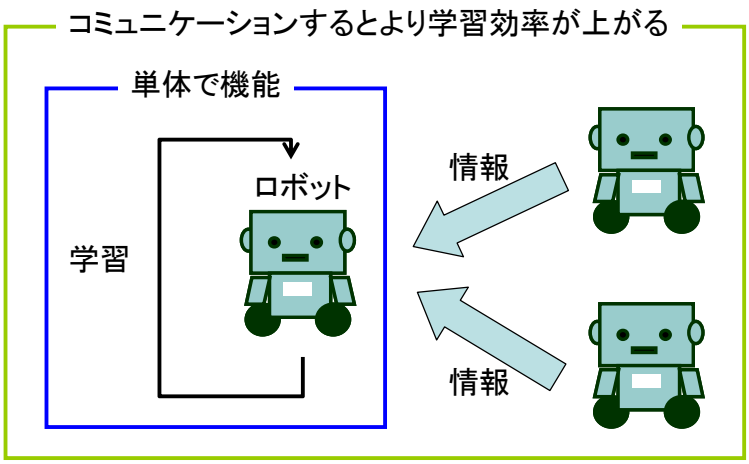


図 2.3　目指すシステム

2.1.2　他者から得られる情報の考察

本システムで重要なのは，コミュニケーションした結果得た他者からの情報（他者情報）の扱い方である．そこで，コミュニケーションによって得られる情報について考察する．

- **扱うことのできる情報について**

  機械で扱う情報には，形式が存在する．情報の形式とは，情報の記述の仕方である．機械では，情報は数値によって表す．例えば，カラー画像情報であれば，ｘ軸とｙ軸からなる２次元配列の要素に色彩を表す値が記述される．また，マイク情報であれば，時間軸の配列に音情報が要素として記述される．

  自身が試行錯誤により得た情報は，自身で扱うことの出来る形式で記述されているため，扱うことができる．他者からの情報は自身が扱える形式で記述されているとは限らない．他者情報が自身で扱えない形式の場合は，その情報を扱うことできない．よって扱うことの出来る他者情報は自身と形式が同一である必要がある．

- **利用出来る情報について**

  情報の形式が自身と同一で，利用可能な情報であればよい．利用可能な情報とは，自身の意思決定に反映できるように記述された情報である．例えば，温度情報の場合を考える．自身の利用可能な温度情報がセ氏で，他者からの情報もセ氏で記述されていれば，その他者からの情報は，自身が利用可能な情報であるといえる．

  他者情報は，形式が自身と同一でも，情報が利用可能な形で記述されていない場合がある．そのような時は，情報を自身が利用可能な形に変換する．例えば，他者情報は絶対温度であり，自身が扱う情報がセ氏であった場合，温度の単位が異なるため，情報をそのまま利用すると悪影響が出る可能性がある．そのため，温度情報を絶対温度からセ氏に変換して利用する．

- **利用の仕方について**

  他者からの情報は，自身の試行錯誤によって得た情報と合わせて意思決定に使われる．自身の得た情報と他者情報を合わせるには，情報を処理する必要がある．情報の処理の仕方は，タスクの種類と目的によって異なる．

  情報の処理の仕方が個体間で異なった場合，コミュニケーションする情報によって評価が異なる場合がある．目的が個体間で異なっている例として，図 2.4 のような迷路タスクで考える．一方は，迷路中に落ちている金を出来るだけたくさん獲得してゴールすることが目的である．もう一方は最短ルートでゴールすることが目的である．両者がコミュニケーションする情報を自身にとって最適のルートの情報とすると，この両者でコミュニケーションした場合，互いに相手の最適なルート情報は，自身にとっては却って悪いルートである．そのため，価値の低い情報となる．このようなことを回避して他者情報を利用するためには，相手のタスク・目的に合わせて，自身の他者情報の処理方法を考えなくてはならない．先の迷路タスクを例にすると，相手の最適なルート情報は，自身にとって通るべきではないルートとして処理することで利用する．

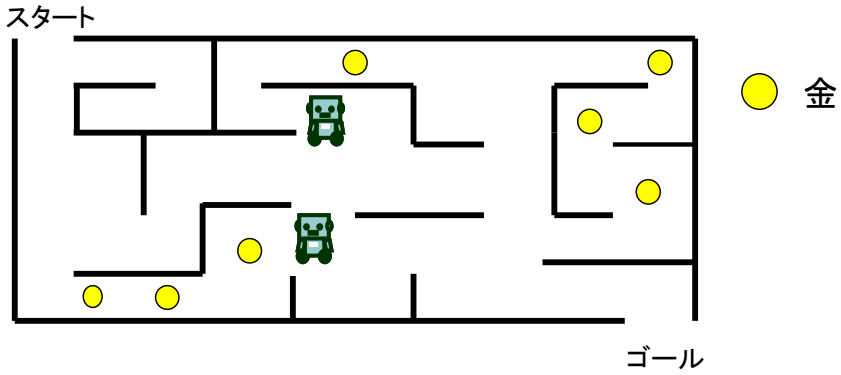


図 2.4　迷路問題

2.2　コミュニケーションについての考察

2.2.1　コミュニケーションの条件

情報の形式については，コミュニケーションをする個体間で同一なのかそうでないのかによって受け取った他者情報を扱うことが出来るかどうかが決まる．他者情報を扱えるようにするためには，情報の形式が個体間で同一である必要がある．

コミュニケーションは他者情報の処理の仕方によって難しさが変わる．利用可能な情報について考えと，他者から得られた情報を自身が利用可能でなければ，自身が利用可能な形に変換必要があるため，その分難しくなる．

情報の処理の仕方については，タスクの種類・目的によって異なってくる．コミュニケーション相手とタスクの種類または目的が異なると，相手ごとに情報の処理方法を考えなければならず，コミュニケーションが難しくなる．逆にタスクの種類・目的が同じであれば，コミュニケーションする相手ごとに情報の処理方法を考える必要がないため，コミュニケーションが簡単になる．

このようにコミュニケーションにも難易度が存在する．そこで，コミュニケーションの中で出来るだけ簡単なコミュニケーションの仕方について考える．まず，コミュニケーションするための前提条件として，以下を考える．

- 情報の形式が個体間で同一であること

また，本研究ではコミュニケーションの一例として，出来るだけ簡単なコミュニケーションを行うために以下を考える．

- 情報の処理方法が個体間で共通であること

そして，上記２つの条件を満たすための個体の十分条件として以下を考える．

- 身体構造が同一であること
- 目的・タスクの種類が同一

身体構造が同一であれば，コミュニケーションによって扱われる情報は，自身の身体構造に準拠したものとなるため，自身で利用することが可能である．目的・タスクが同一であれば，やり取りされる情報は同じ目的・タスクに関するものなので利用しやすくなる．

2.2.2　コミュニケーションする情報

2.2.1 において考えた簡単なコミュニケーションをするための個体の十分条件は，身体構造が同一であること，目的・タスクの種類が共通であることであった．これらの条件をもとに本論文においてコミュニケーションする情報を考える．

コミュニケーションする情報は，それぞれの個体への依存度が低い方がよい．個体に依存する情報は，状況に関する情報と身体に関する情報である．状況に関する情報とは，個体自身がセンサー等で感知することの出来る周囲の環境の状態である．身体に関する情報は，自身の身体そのものに関する情報である．身体構造は個体の身体の構成要素（人間であれば頭，胴体，腕・足が共に２本）であるのに対し，身体は，人間で言うなら背の高低や体格の大小といった情報のことを指す．状況や身体に関する情報はその個体固有の情報である．そういった情報はその個体であるから利用しやすいのであって，他者から見れば利用が困難な場合がある．よって，状況や身体といった個体への依存度が高い情報はコミュニケーションに用いず，個体への依存度が低い情報をコミュニケーションに用いる．そこで，本論文ではそういった情報に「学習法」を選択することにする．

2.3　学習法をコミュニケーション情報とした個体知能の発達

学習法をコミュニケーション情報とした個体知能の発達方法の概念図を図 2.5 に示す．この概念図より，個体の学習は，行動学習部と学習法学習部の２つの部分で構成する．
学習全体の流れを以下に示す．

1．学習法学習部で自身の知識に基づいて学習法を決定する．
2．決定した学習法を適用し，行動学習部で行動の選択が行われる．
3．選択した行動を実行し，結果を得る．
4．結果から，学習法決定部では決定した学習法を評価し，行動学習部では，選択した行動を評価する．学習法決定部，行動学習部での評価が自身の知識となる．
5．コミュニケーションによる他者情報を評価し，自身の知識とする．

１～５を繰り返すことで学習を行う．

次に行動学習部と学習法学習部の学習の流れをそれぞれ見ていく．

- **行動学習部**

  行動学習部では，学習法学習部で決定した学習法を用いて，個々の状況（その時その時の周囲の状態）に合った行動を学習する．行動学習は以下の流れで行われる．

  １．学習法に基づいて自身の知識から，どのような行動を実行するか決定する．
  ２．行動を実行した結果が出る．
  ３．学習法に基づいて結果を評価し，自身の知識とする．

  １～３を繰り返すことで直面する環境に適した学習を行う．

- **学習法学習部**

  学習法学習部では，自身の直面する環境に適した学習法を学習する．学習は以下の流れで行われる．コミュニケーションによる他者情報は，他者が行った学習法とその結果である．

  １．現在直面している環境と自身の知識から，学習法を決定する．
  ２．決定した学習法を用い，行動学習部で行動を学習する．
  ３．行動学習の結果が出る．
  ４．その結果から決定した学習法を評価し，自身の知識とする．
  ５．コミュニケーションにより得られる他者情報を評価し，自身の知識とする．

  １～５を繰り返すことで，自身の直面している環境に合った学習法を学習し，決定する．

学習法学習部と行動学習部の関係を図 2.6 に示す. 学習法学習部ではどのような学習法がよいかを学習し，行動学習部では学習法学習部で決定した学習法を用いてどの行動がよいかを学習する．

第４章では，この概念を強化学習の枠組みに適用したシステムを提案し，第５章で有効性の検証を行う．

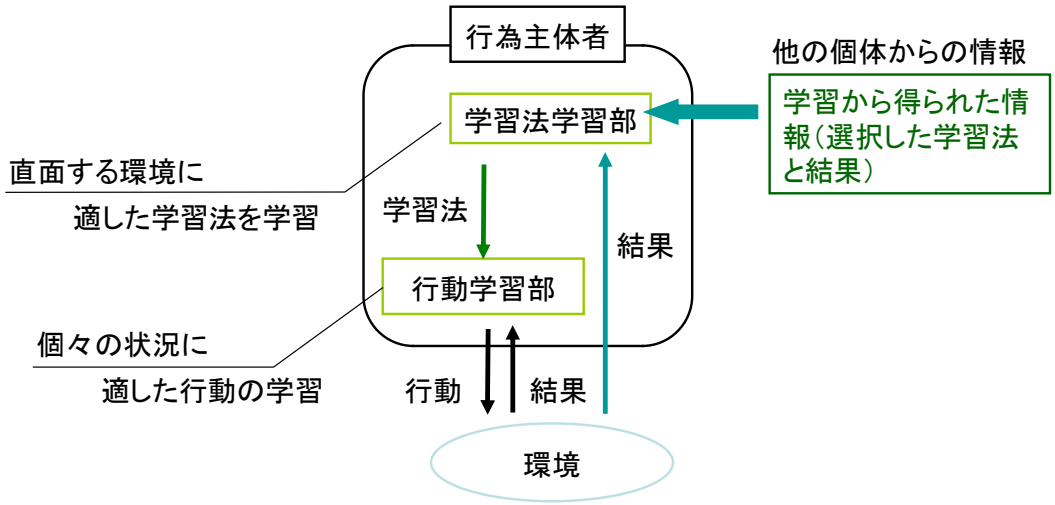


図 2.5　学習法をコミュニケーション情報とした個体知能の発達方法の概念図

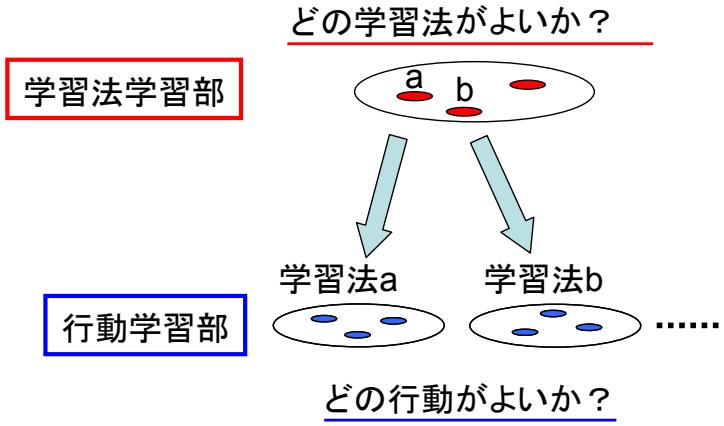


図 2.6　学習法学習部と行動学習部の関係

第３章　強化学習

本章では強化学習[2]について述べる．3.1 では強化学習の概要を述べる．3.2 では，行動選択手法について代表的なものを述べる．3.3 では行動評価手法について本論文で扱うものについて述べる．

3.1　強化学習の概要

強化学習は，環境との試行錯誤による相互作用を通して適切な行動戦略を獲得する機械学習である．

3.1.1　強化学習の特徴

強化学習の特徴としては以下のようなことが挙げられる．

- 試行錯誤的な探索
- 遅延報酬

試行錯誤的な探索

強化学習では，学習のための正解情報を直接与えられることはなく（教師なし学習），実行した行動の評価を訓練情報として利用する．従って評価の高い行動を直接探索するための試行錯誤による能動的な探索が必要となる．実行した行動がどれくらい良いものかが知らされ，それが可能な行動の中で最良または最悪であるかについては知らされない．

遅延報酬

強化学習では，行った行動に対してその行動の良し悪しを報酬という数値を用いて表す．エージェントは報酬を基に行動を評価し，その評価に従い行動を学習する．遅延報酬は，行動は直接的な報酬のみならず，その次の状況に影響を与え，そのことを通じて，その後に続く全ての報酬に影響を与えるという性質のことである．

3.1.2　環境とエージェントとの相互作用とエージェントの目的

強化学習における行為者（エージェント）と環境との相互作用の概念図を図 3.1 に示す．エージェントは環境に対して，以下のような状況の観測，行動，報酬の獲得という一連の作業を行う．

1．エージェントは時刻 t で環境から知覚される状態 s_t に基づいて意思決定を行い，行動 a_t を取る．

２．エージェントの行動 a_t により，環境は s_{t+1} へ遷移する．
３．エージェントの行動 a_t の結果として環境から報酬 r_t を受け取る．

エージェントの目的は環境との相互作用によって自身が獲得する報酬和を最大化することである．

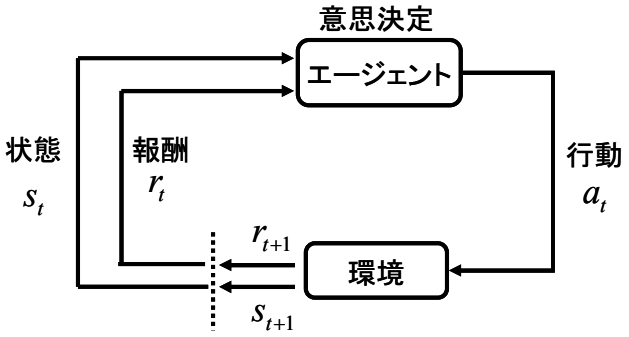


図 3.1　環境との相互作用

3.1.3　強化学習の構成要素

ここでは強化学習の構成要素である，方策（行動選択手法）・報酬関数・価値関数（行動価値の評価・推定）について解説する．

・方策（行動選択手法）

方策はある時点での学習エージェントのふるまい方を定義する．方策は，環境において知覚した状態から，その状態にあるときに取るべき行動への写像である．この方策は，一般的には確率的である．

・報酬関数

報酬関数は，環境において知覚した状態 s でとった行動 a（状態行動対）の報酬を決定する．この報酬はその状態 s で行動 a をとる望ましさを表している．強化学習エージェントの唯一の目的は，最終的に受け取る報酬を最大化することである．この報酬関数はエージェントにとって何が良い出来事で何が悪い出来事であるかを定義している．報酬関数はエージェントが変更できないものである．しかし，方策を変更する際の判断材料として使うことができる．一般的には報酬関数は確率的である．

・価値関数

報酬関数が即時的な意味合いで何が良いのか示しているのに対して，価値関数は，最終的な状態または行動の価値を決定する．価値とは，エージェントがその

状態を基点として将来にわたって入手できる報酬の期待値である．報酬はその環境が即時的で固有の望ましさを決定するのに対して，価値はその後に続きそうな状態群とそれらの状態群で得られそうな報酬を考慮に入れた上での長期的な望ましさを示すものである．

3.1.4　強化学習の流れ

強化学習は，3.1.2 で示した環境との相互作用により行われる．強化学習の流れを図 3.2 に示す．環境から知覚した状態 s によって，エージェントは自身が行うことのできる行動の中から，その状態における行動価値に基づき行動選択手法を用いて行動 a を選択（意思決定）し，実行する．その結果，環境より得られた報酬 r を基に，エージェントは状態 s において選択した行動 a の価値 $Q(s,a)$ の更新を行動評価手法によって行い（学習），次回同様の状態においての行動選択に生かす．この流れより，強化学習の学習法は**行動選択手法と行動評価手法の 2 つの組み合わせ**によって決定する．

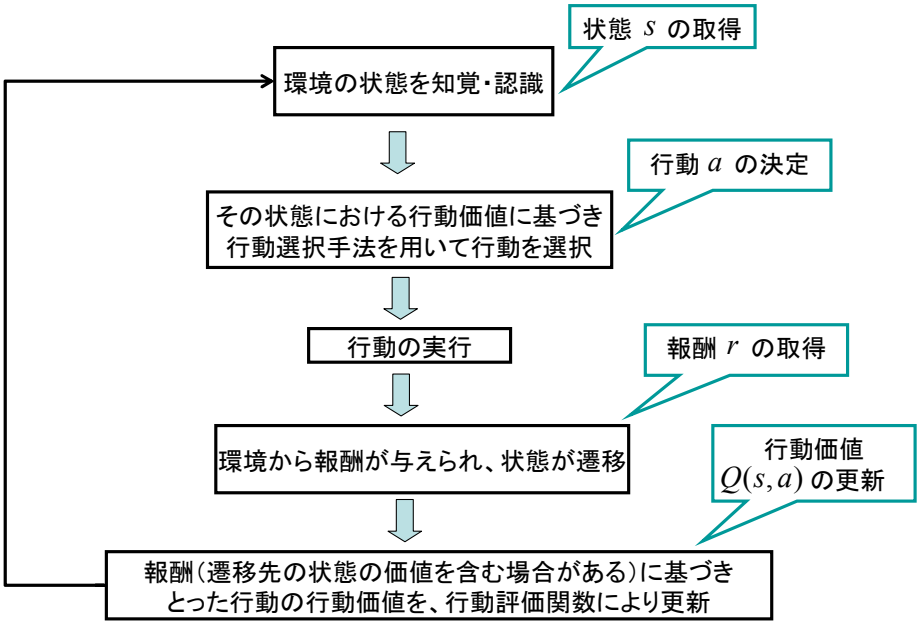


図 3.2　強化学習の流れ

3.2　行動選択手法

行動選択手法とは，エージェントが認識した状態 s においてとる行動 a を選択する際に用いられる手法である．ここでは，代表的な行動選択手法として，greedy 法，ε - greedy 法，softmax 法，追跡手法について述べる．

3.2.1　greedy 法

直面する状態において，最も高いと推定された行動価値を持つ行動（あるいは行動群から１つ）を選択する（図 3.3）．この方法は常に即時の報酬を最大にするために，現在の知識を利用するものである．すなわち，価値が低いと判断される行動に対しては，その行動の価値が一時的に低いだけで，本当は価値の高い行動であるという可能性を確かめるための試行を一切行わない．そのため，真に価値の高い行動を選択しにくい．

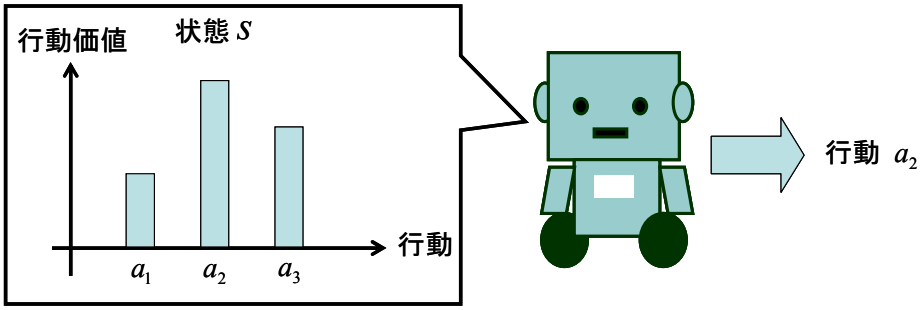


図 3.3　greedy 法

3.2.2　ε - greedy 法

直面する状態において，基本的には推定される行動価値が最も高い行動（グリーディな行動）を選択するが，たまに小さい確率 ε で行動価値の高さとは無関係にランダムで行動を選択する手法である．常に知識利用しか行わない greedy 法とは異なり，確率 ε で探査行動を行う．しかし，確率 ε における行動選択の際にほとんど最悪と思われる行動とほとんど最適に近い行動を選択する可能性が同じくらいの高さになるという欠点がある．

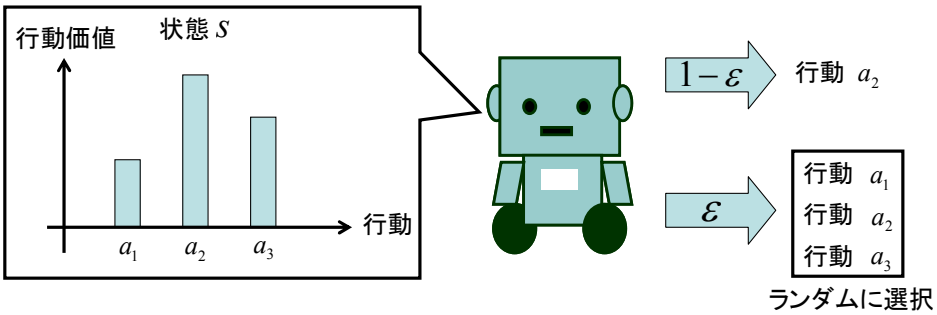


図 3.4　ε - greedy 法

3.2.3　softmax 法

softmax 法は，行動価値に基づいて行動確率を変化するやり方である．すなわち，行動価値の最も高い行動には最も高い選択確率が与えられ，他のすべての行動は，その行動価値

に従って重みをかけられ，ランク付けされる（図 3.5）．

Softmax 法では一般に，Gibbs 分布，あるいは Boltzmann 分布が使われる．具体的には，時間 t における状態 s で行動 a を選択する確率 $\pi_t(s,a)$ は式(3.1)で与えられる．

$$\pi_t(s,a)=\frac{e^{Q_t(s,a)/\tau}}{\sum_{b=1}^{n} e^{Q_t(s,a)/\tau}} \qquad (3.1)$$

$\pi_t(s,a)$：時間 t，状態 s で行動 a を選択する確率

$Q_t(s,a)$：時間 t，状態 s で行動 a を実行したときの行動価値

τ ：温度

ここで τ は温度と呼ばれる正定数である．τ が高い場合には，すべての行動が（ほぼ）同程度に起こるように設定される．また，τ が低い場合には，行動価値の異なる動作において選択確率の差がより大きくなるように設定される．そして，$\tau \to 0$ の極限では，softmax 法は greedy 法と一致する．

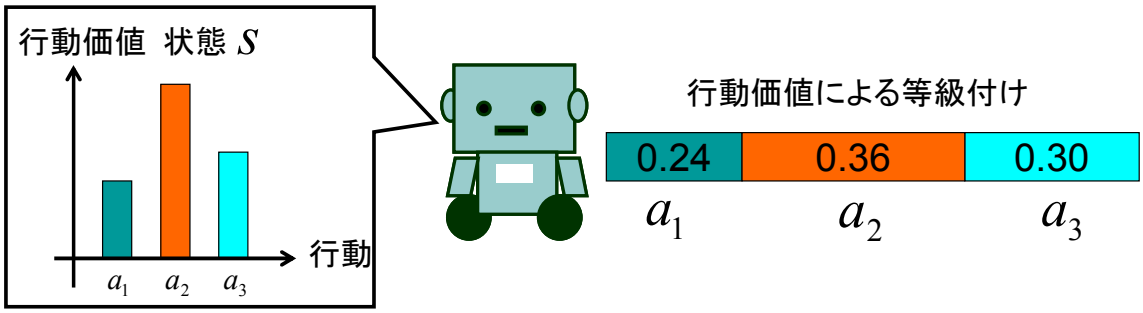


図 3.5　softmax 法

3.2.4　追跡手法

追跡手法では行動価値と行動優先度という値を用いる．行動優先度は現在の行動価値の最も高い行動の選択確率を増加させる目的で使われる．最も単純な追跡手法では，行動優先度として状態 s において時間 t で行動 a を選択する確率 $\pi_t(s,a)$ を用いる．

毎回の行動の直後，最も行動価値の高い行動が選ばれる可能性がより高くなるように，この確率値が変更される．状態 s において時間 t での行動の後，時間 $t+1$ に対する最も行動価値の高い行動（複数個ある場合には，その中からランダムに 1 つ）が $a^*_{t+1}=\arg\max_a Q_{t+1}(s,a)$ であるとする．この場合，行動 $a_{t+1}=a^*_{t+1}$ の選択確率は，式(3.2)のように，確率 1 に向かって β の比率で増加させられる．

$$\pi_{t+1}(a^*_{t+1})=\pi_t(a^*_{t+1})+\beta\left[1-\pi_t(a^*_{t+1})\right] \qquad (3.2)$$

残りの行動の選択確率は，式(3.3)のように，すべての $a \neq a_{t+1}^{*}$ に対して 0 に向かって減らされる．

$$\pi_{t+1}(a_{t+1}^{*}) = \pi_t(a_{t+1}^{*}) + \beta[0 - \pi_t(a)] \qquad (3.3)$$

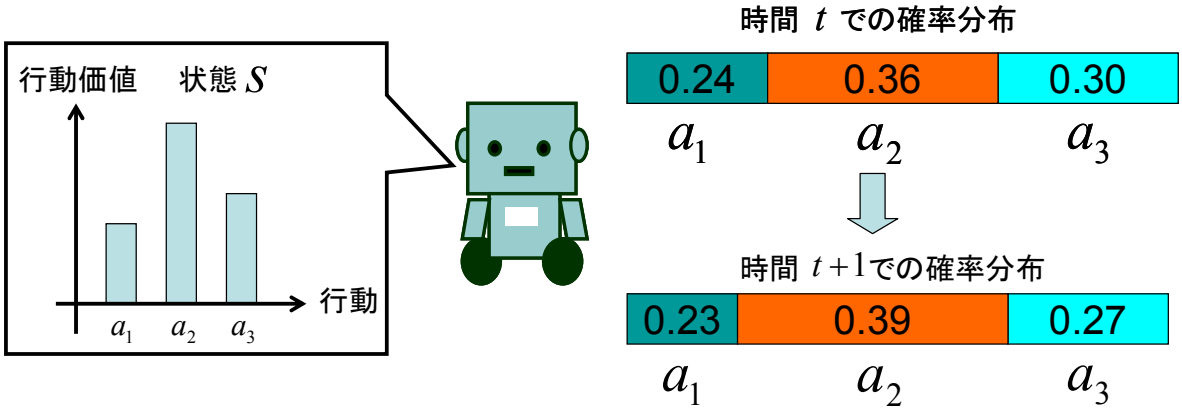


図 3.6　追跡手法

3.2.5　強化比較法

強化学習は，大きい報酬に続く行動はより再現しやすく，小さい報酬に続く行動はより再現しにくくなるという経験的事実を基本とした方法である．強化比較法は，得られた報酬が大きい報酬なのか，小さい報酬なのかを判断するために基準となる報酬（リファレンス報酬）を用いる手法である．リファレンス報酬としては，今まで受け取った報酬の平均であることが多い．リファレンス報酬より大きい報酬を得られた行動は，大きい報酬と判断され，リファレンス報酬よりも小さい報酬が得られた行動は，小さい報酬であると判断される．

強化比較法は行動価値推定量を持たないため，行動評価手法と組み合わせて使用しない，少し特殊な手法である．強化比較法では行動価値推定量の代わりに，全体的な報酬レベルを表すリファレンス報酬を更新する．

行動優先度の更新については，得られた報酬を，リファレンス報酬を $\bar{r}_t$ とすると行動優先度は式(3.4)で更新される．

$$p_{t+1}(s,a) = p_t(s,a) + \beta[r_t - \bar{r}_t] \qquad (3.4)$$

$p_t(s,a)$ ：時間 t，状態 s において行動 a を選択する優先度

r_t ：得られた報酬

$\bar{r}_t$ ：リファレンス報酬

β ：ステップサイズ・パラメータ（$0 \le \beta \le 1$）

βは正のステップサイズ・パラメータを示す．この式は，リファレンス報酬と比べ，高い報酬を得た場合にはその行動を再選択する確率を増やし，低い報酬の場合には再選択確率を減少させる．

行動の選択は，通常 softmax 手法（3.2.3 参照）が用いられる．確率決定を式(3.5)に示す．

$$\pi_{t+1}(s,a) = \frac{e^{p_{t+1}(s,a)}}{\sum_{b=1}^{n} e^{p_{t+1}(s,b)}} \qquad (3.5)$$

$\pi_t(s,a)$：時間t，状態sにおいて行動aを選択する優先度

リファレンス報酬は，選択した行動にかかわらず，最近受け取った報酬すべてを逐次平均して計算される（式(3.6)）．

$$\bar{r}_{t+1} = \bar{r}_t + \alpha\left[r_t - \bar{r}_t\right] \qquad (3.6)$$

α ：ステップサイズ・パラメータ（$0 \le \alpha \le 1$）

3.3 行動評価手法

強化学習において，エージェントは行動の真の価値そのものを知ることはできないため，毎回の行動によって得られる報酬からその行動の真の価値を推定する．そして，その推定値を使って行動選択手法を通して行動を選択する．この行動の真の価値を推定するための方法が行動評価手法である．

ここでは，行動評価手法として，標本平均手法，加重平均手法，Q 学習法の 3 手法について述べる．

3.3.1 標本平均手法

標本平均化手法では，その行動が選ばれたとき実際に得られた報酬を平均化してゆく．状態s，行動aについて標本平均手法を用いたときの時間tでの行動の価値$Q_t(s,a)$は式(3.7)のように表される．

$$Q_t(s,a) = \frac{r_{s1} + r_{s2} + r_{s3} + \cdots + r_{sk_{sa}}}{k_{s,a}} \qquad (3.7)$$

$Q_t(s,a)$：時間t，状態sで行動aを選択したときの行動価値

k_{sa} ：状態sで行動aの累計選択回数

$r_{s1}+r_{s2}+r_{s3}+\cdots+r_{sk_{sa}}$ ：状態 s で行動 a が選択されたそれぞれの時間の獲得報酬

$k_{sa}=0$ の場合には，$Q_t(s,a)$ を，$Q_0(s,a)=0$ のような初期値にする．大数の法則より，$k_a \to \infty$ の極限において，$Q_t(s,a)$ は真の価値 $Q^*(s,a)$ に収束する．標本平均化手法は定常環境での動作に適したものである．

3.3.2 加重平均手法

加重平均手法は，遠い過去の報酬よりも最近に受け取った報酬の方により重みを与えるような方法である．重みを与えるために，定数値のステップサイズ・パラメータを使用する．行動 a をとったときの行動価値値 $Q_t(s,a)$ を更新するための更新式は式(3.8)のようになる．

$$Q_{t+1}(s,a)=Q_t(s,a)+\alpha\left[r_{t+1}-Q_t(s,a)\right] \tag{3.8}$$

$Q_t(s,a)$：時間 t，状態 s で行動 a を選択したときの行動価値

r_{t+1} ：新たに獲得した報酬

α ：ステップサイズ・パラメータ（$0 \le \alpha \le 1$）

3.3.2 Q 学習法

Q 学習は，現在の状態で選択した行動の価値とその行動の結果遷移した先の状態の行動価値によって，現在の行動価値を更新する（図 3.8）．Q 学習での行動価値更新式は式(3.10)のようになる．γ は割引率である．割引率は次の状態の行動価値が現在においてどれだけの価値があるかを決定する．γ が小さければ，現在の行動価値を重視し，γ が大きければ，遷移先の行動価値を重視する．

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t)+\alpha\left[r_t+\gamma m\underset{a}{a}xQ(s_{t+1},a)-Q(s_t,a_t)\right] \tag{3.10}$$

$Q(s_t,a_t)$：時間 t，状態 s_t で行動 a_t をとった時の行動価値

r_t ：報酬

α ：ステップサイズ・パラメータ（$0 \le \alpha \le 1$）

γ ：割引率（$0 \le \gamma \le 1$）

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left[r_t + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)\right]$$

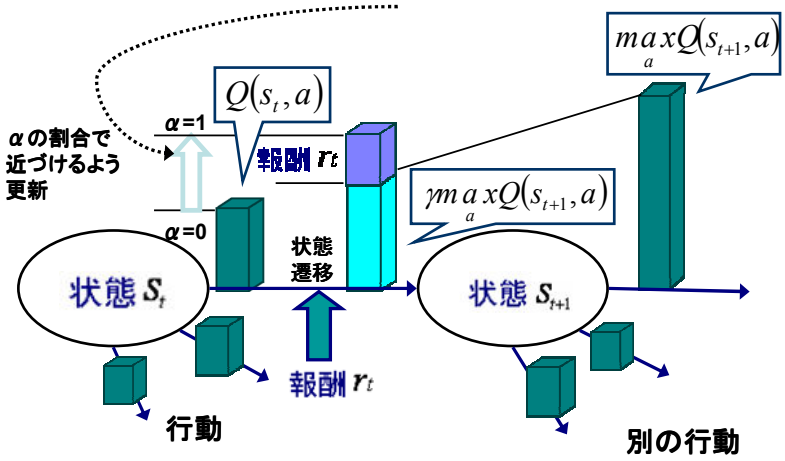


図 3.8　Q 学習概念図[1]

3.4　まとめ

本章では，本論文で扱う学習法である強化学習について説明した．3.1 では強化学習の概要について解説し，3.2 で行動選択手法，3.3 で行動評価手法について説明した．

第４章　強化学習を適用したコミュニケーションによる学習システム

4.1　学習法をコミュニケーションする個体知能の発達概念

図 4.1 に学習法をコミュニケーション情報とした個体知能の発達方法の概念図を示す．この概念図より，個体の学習は，学習法学習部と行動学習部に分かれている．学習法学習部ではどのような学習法がよいかを学習し，行動学習部では学習法学習部で決定した学習法を用いてどの行動がよいかを学習する．学習全体の流れを以下に示す．

１．学習法学習部で自身の知識に基づいて学習法を決定する．
２．決定した学習法を適用し，行動学習部で行動の選択が行われる．
３．選択した行動を実行し，結果を得る．
４．結果から，学習法決定部では決定した学習法を評価し，行動学習部では，選択した行動を評価する．学習法決定部，行動学習部での評価が自身の知識となる．
５．コミュニケーションによる他者情報を評価し，自身の知識とする．

１～５を繰り返すことで学習を行う．

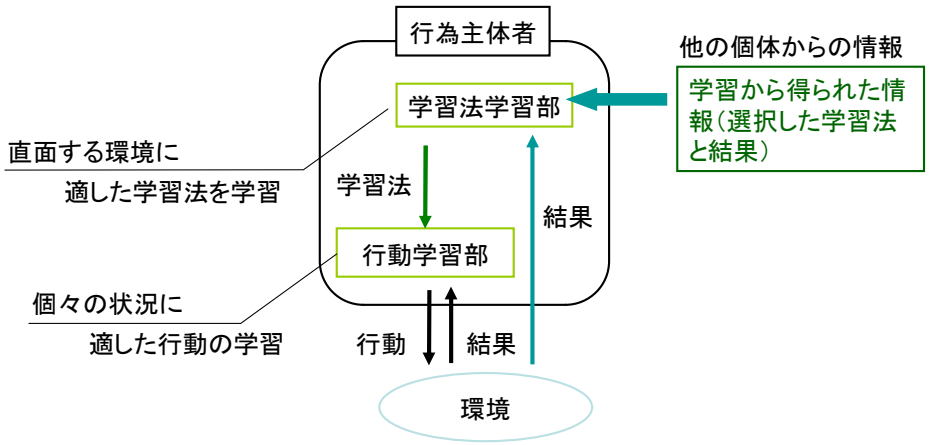


図 4.1　学習法をコミュニケーション情報とした個体知能の発達方法の概念図

4.2　強化学習を適用したコミュニケーション学習システム

4.1.1 で説明した概念に強化学習を適用した提案システムを図 4.2 に示す．強化学習を適用したことにより明確になった部分について述べる．

- **学習法について**

  強化学習を適用したシステムにおける学習法は，行動選択手法＋行動評価手法の組み合わせとなる．例えば第３章で説明した行動選択手法と行動評価手法で考えるならば，ε - greedy 法（行動選択手法）と標本平均手法（行動評価手法）となる．なお，強化比較法については，行動評価手法を必要としないため，それ単体で１つの学習法とする．

- **結果について**

  図 4.1 の概念図にある結果は，強化学習適用システムにおいて報酬となる．

- **学習法学習部**

  図 4.2 より学習法学習部は行動選択と行動評価を行う．学習法学習部における行動選択は学習法の選択に相当する．行動選択・行動評価には，第３章で述べた行動選択手法・行動評価手法が用いられる．学習法学習部の行動評価手法は，自己の報酬を評価するものと，他者情報を評価するものの２つが存在する（図 4.3）．

- **行動学習部**

  図 4.2 の行動学習部の行動選択・行動評価には，学習法学習部で決定された学習法に設定された行動選択手法と行動評価手法が適用される．

- **他者情報について**

  他者情報としてやり取りされるものを以下に示す

  １．他者の直面している環境
  ２．他者の採用する学習法
  ３．他者の得た報酬

・行動学習部の流れ

学習の流れを図 4.4 に示す．現在の環境状態を認識し，その環境状態に対する学習法を行動選択手法により選択する．選択した学習法を実行し，報酬を得る．そして，選択した行動により得られた報酬からその学習法に対する評価を行動評価手法により更新する．行動学習部の流れは，図 4.5 の学習法の実行の部分に当たる．

・学習法決定部の学習の流れ

学習の流れを図 4.5 に示す．現在の環境状態を認識し，その環境状態に対する学習法を行動選択手法により選択する．選択した学習法を実行し，報酬を得る．そして，選択した学習法で得られた報酬により，その学習法に対する評価を行動評価手法により更新する．また，他者情報から行動評価手法を用いて自身の学習法の価値を更新する．

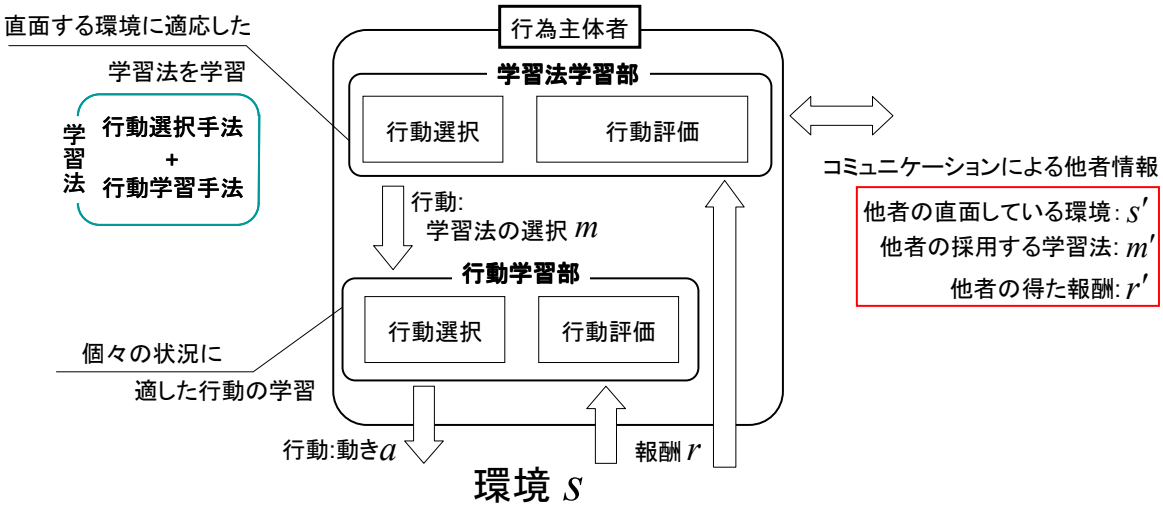


図 4.2　強化学習を適用したコミュニケーション学習システム

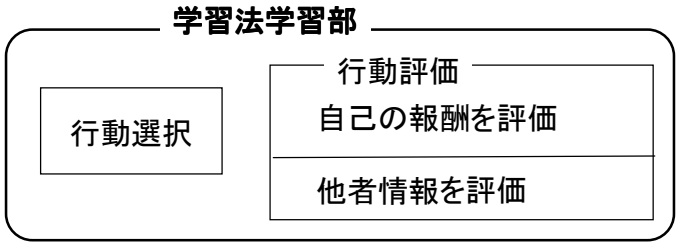


図 4.3　学習法学習部の評価

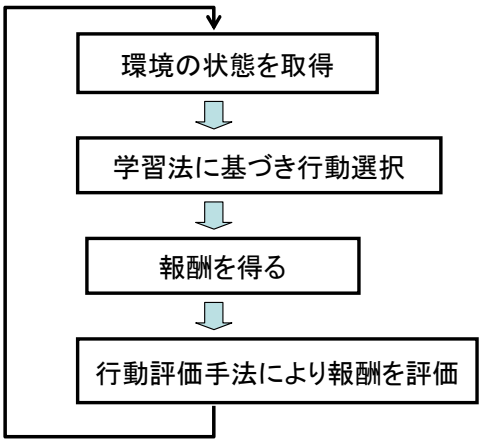


図 4.4　行動学習部の学習の流れ

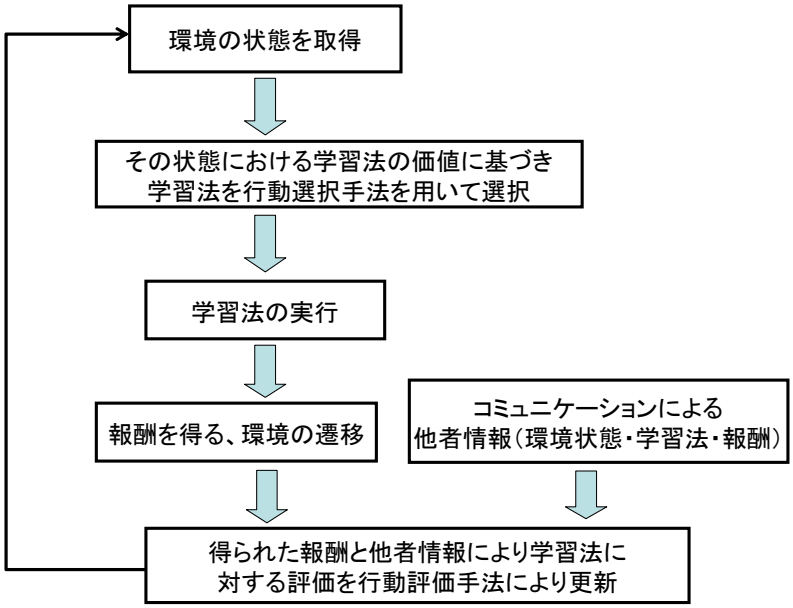


図 4.5　学習法決定部の学習の流れ

4.3　まとめ

本章では，2.2 の概念図をもとに，提案システムとして強化学習を適用したコミュニケーションによる学習システムを構築し，システムの概念について述べた.

第５章　N 本腕バンディット問題を対象とした提案システムの有効性の検証

5.1　N 本腕バンディット問題とは

N 本腕バンディットマシンとは，N 本の腕（レバー）のあるスロットマシンのことである．スロットマシンの各腕にはそれぞれ当選確率が設定されており，当選すると報酬が支払われる（図 5.1）．N 本腕バンディット問題は，N本腕のバンディットマシンに対して，得られる報酬を最大にすることが目的である．

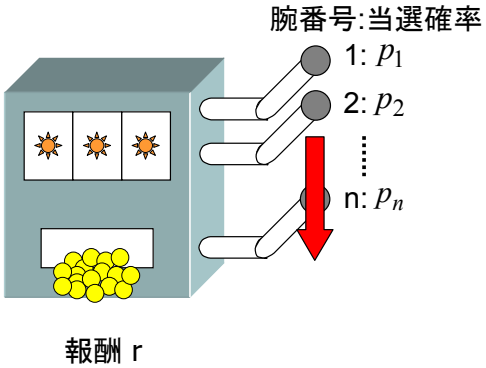


図 5.1　N 本腕バンディット

5.2　実験概要

本実験では，N本腕バンディット問題に対して，強化学習を適用したコミュニケーションを用いた学習システム（以降，提案システム）の有効性を検証する．実験環境として，各バンディットの腕の当たり確率が試行の度に変化するような非定常環境を考える．そのような環境に対してバンディットマシンは，自身の試行錯誤による学習とコミュニケーションによる他者情報を使い，直面する環境に合った学習法と直面する状況に合った腕の選択を学習する．

・実験目的

それぞれのエージェントが直面している環境について適した学習法を選択していることを確認する．

・実験環境

実環境では他者と同じ環境にあるということはないため，各エージェントはそれぞれ異なる環境に直面することを考える．エージェントはそれぞれの直面する環境に合

った学習法と状況に合った行動（最も当たり確率の高い腕）を学習する．そのために，本実験ではバンディットマシンの腕の当選確率を試行毎に変動するような実験環境を考える．変動の仕方（変動のしやすさ，変動の大きさ）はバンディットマシンによってそれぞれ異なるようにする．こうすることによってそれぞれ異なった環境を構築することができる．

腕の当選確率の変動の仕方は，変動頻度（Th）と変動振幅（Amp）の２つの指標によって決定することを考える．変動頻度は試行毎に当選確率の変動が起きるかどうかを確率的に決定するための変数で，$0 \le Th \le 1$ の範囲の値をとる．変動頻度が高いと試行毎に確率変動が起きる場合が多くなり，低いと少なくなる．また，当選確率の変動値は変動振幅に従って式 5.1 で決定される．

$$\alpha_i(n) = \alpha_i(n-1) + RAND \times Amp \times 2 - Amp \qquad (5.1)$$

$\alpha_i(n)$ ：n 回目試行でのバンディットの i 番目の腕の当たり確率
$RAND$ ： $0 \le RAND \le 1$の範囲でランダムに決定した実数
Amp ：変動振幅（$0 \le Amp \le 1$）

変動振幅は腕の当選確率の変動量の絶対値である．例えば，$Amp = 0.5$ であれば，$-0.5 \sim 0.5$ の範囲で腕の当選確率が変動する．変動振幅の範囲内でランダムに決定された値を前回の試行での腕の当たり確率に加えることで，新たな腕の当たり確率とする．

エージェントが直面する環境マップの設定として，変動振幅・変動頻度を用いたマップを構築する（図 5.2）．このマップのそれぞれのマスがバンディットマシンにあたる．配置された１台のバンディットマシンに１体のエージェントが対応する（図 5.3）．このような環境マップは，変動振幅，変動頻度が少しずつ変わっていくため，周辺のエージェントは自身と似たような環境にある．したがって，コミュニケーションが行いやすい．

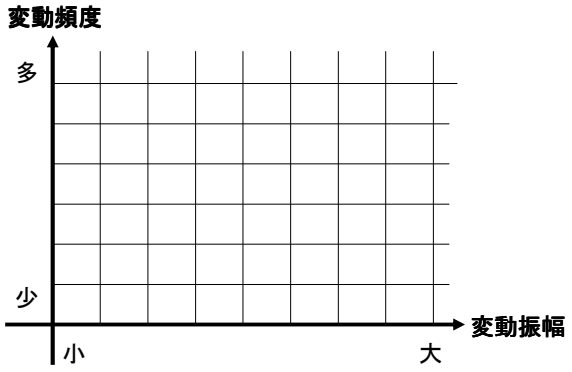


図 5.2　バンディットマシン環境マップ

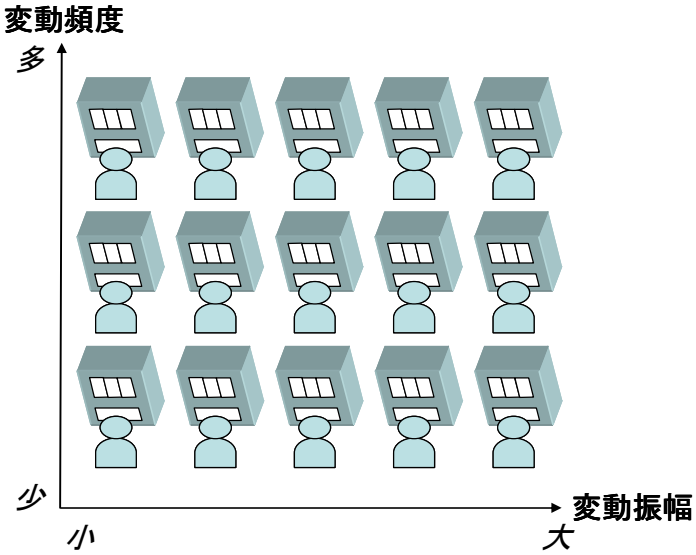


図 5.3　バンディットとエージェントの対応

次に変動頻度の多少，変動振幅の大小による環境の特徴について説明する．説明する環境は，以下の４つの環境である．

- ■　変動頻度：少　変動振幅：小　（図 5.4(a)）
- ■　変動頻度：少　変動振幅：大　（図 5.4(b)）
- ■　変動頻度：多　変動振幅：小　（図 5.4(c)）
- ■　変動頻度：多　変動振幅：大　（図 5.4(d)）

・変動頻度：少　変動振幅：小（図 5.4(a)）

この環境では，変動頻度が少ないため当選確率はあまり変わらない．また，変動振幅が小さいため，試行開始時に設定した当選確率から大きく変化することがない．基本的に変化の少ない静的な環境といえる．

・変動頻度：少　変動振幅：大（図 5.4(b)）

この環境では，変動頻度が少ないため当選確率はあまり変わらない．しかし，変動振幅が大きいため，当選確率の変動幅が広く，当選確率が大きく変動することがある．

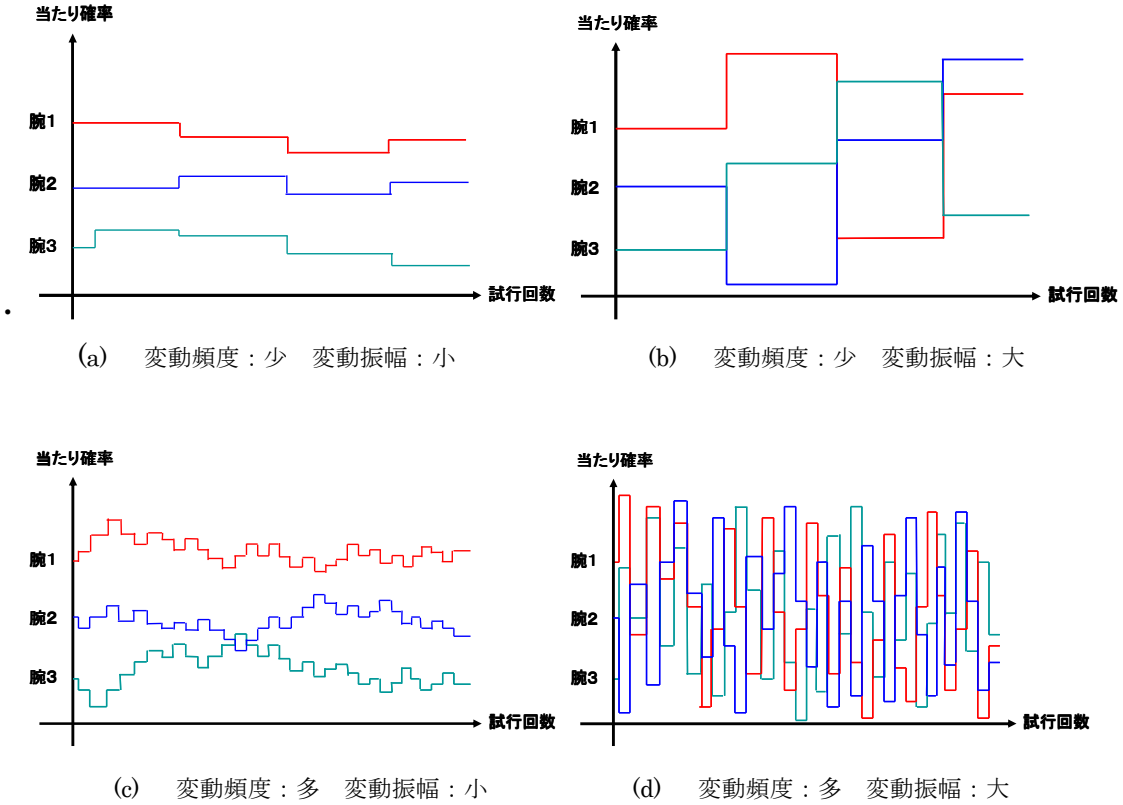
・変動頻度：大　変動振幅：小（図 5.4(c)）

この環境では，変動頻度が多いため当選確率が頻繁に変化する．また，変動振幅が小さいことから，変化の度合いは小さく，試行開始時に設定した当選確率から大

きく変化することがない.

・**変動頻度：大　変動振幅　：大**（図 5.4(d)）

この環境では，変動頻度の値が多いため当選確率が頻繁に変化する．また，変動振幅も大きいことから当選確率の変動幅も広い．したがって，試行するたびに最も高い当選確率を持つ腕が変わるという予測不可能な環境になる．



(a)　変動頻度：少　変動振幅：小　　(b)　変動頻度：少　変動振幅：大

(c)　変動頻度：多　変動振幅：小　　(d)　変動頻度：多　変動振幅：大

図 5.4　変動振幅と変動頻度の違いによるバンディットの確率変動の違い

5.3　実験設定

◆　タスク環境に関する設定

腕の確率変動を決定する変動頻度（Th）と変動振幅（Amp）を設定する環境マップのイメージを図 5.5 に示す．変動頻度と変動振幅を軸とし，Th，Amp は 0 から 1 までマス毎に 0.01 刻みで設定する．よって，10000 種類タスク環境の異なるバンディットマシンができる．

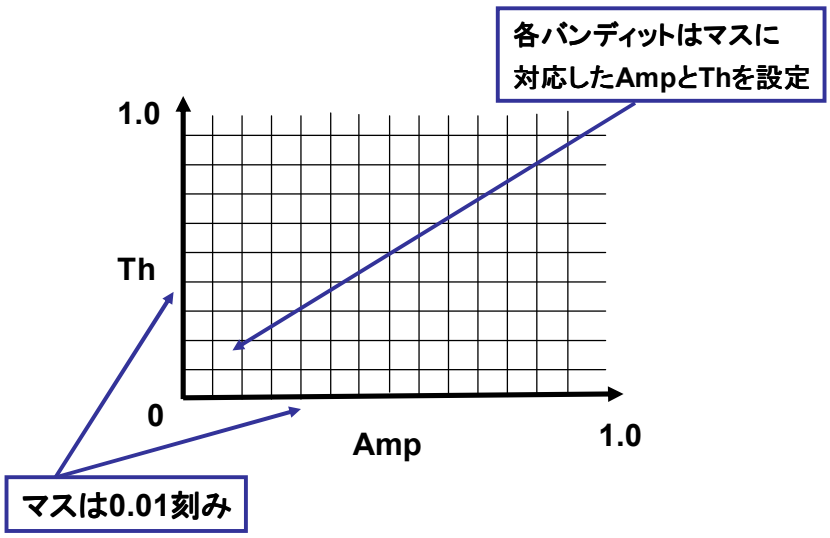


図 5.5　実験環境マップ

◆　バンディットマシンに関する設定

実験に使用するバンディットマシンの台数，バンディットマシンの腕本数，また支払われる報酬については表 5.1 に示す．

表 5.1　バンディットマシンの設定

バンディット台数	10000 台
腕本数	6 本
報酬	1

また，各腕の初期確率について表 5.2 に示す．これは，全てのバンディットマシン共通の設定である．

表 5.2　バンディットマシンの各腕の初期確率

腕番号	1	2	3	4	5	6
確率	0.8	0.61	0.33	0.1	0.01	0.56

◆　エージェントに関する設定

エージェントの数，総試行回数，コミュニケーションに関する設定を表 5.2 に示す．エージェント数は本実験で用いるエージェントの数である．バンディットマシンと 1 対 1 対応させる．総試行回数は，エージェント 1 体当たり何回試行するかという設定である．コミュニケーション頻度は何回試行毎に行うかの設定である．コミュニケーション対象はどの情報の送信・受信対象の設定である．コミュニケーションする情報は相手に送る・相手から受け取る情報の内容である．

表 5.3　エージェント設定

エージェント数	10000 体
総試行回数	30000 回
コミュニケーション頻度	1 回試行毎
コミュニケーション対象	自己の周囲 8 マスに存在するエージェント
コミュニケーションする情報	コミュニケーション時に自己が 適用していた手法と得られた報酬

（学習法学習部，行動学習部に関する設定）

図 5.6 に提案システムの概要図を示す.

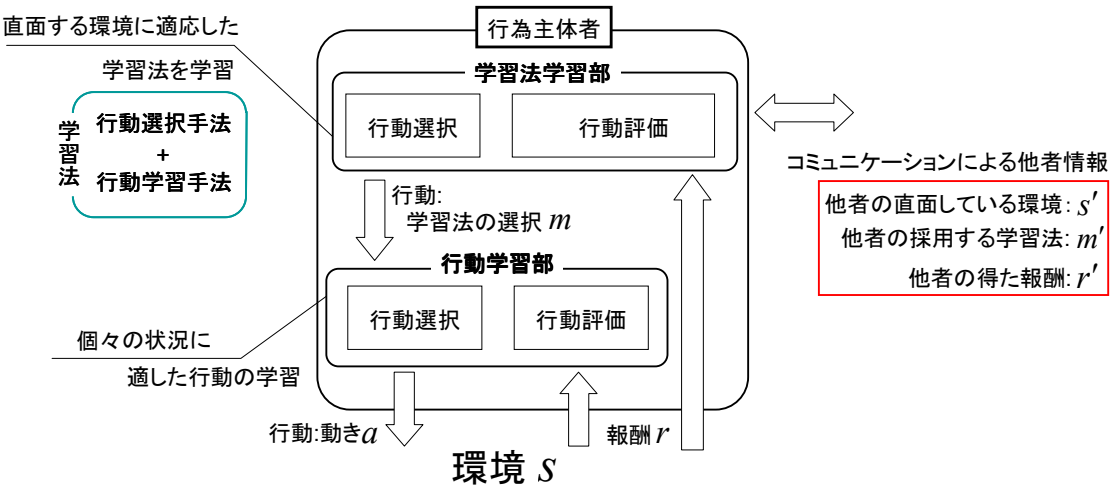


図 5.5　提案システム概要図

・行動学習部で使う学習法

学習法実行部で使う学習法を表 5.4 に示す．ε - greedy 法，softmax 法，追跡手法の 3 つの行動選択手法に，標本平均手法，加重平均手法，Q 学習法の 3 つの学習手法を組み合わせた 9 つの学習法と，強化比較法を加えた計 10 種類の学習法を扱う．これら 10 種類の学習法の中から，自身が直面する環境に合った学習法を学習する．

表 5.4　学習法実行部で使う学習法

行動選択手法	行動評価手法
ε －greedy 法	標本平均手法
	加重平均手法
	Q 学習法
softmax 法	標本平均手法
	加重平均手法
	Q 学習法
追跡手法	標本平均手法
	加重平均手法
	Q 学習法
強化比較法	使用しない

・学習法学習部で使う学習法

学習法を選択する手法と，学習法を評価する手法を表 5.5 に示す.

表 5.5　学習法決定部で使う学習法

行動選択手法	行動評価手法
ε －greedy 法	加重平均手法

今回扱う N 本腕バンディット問題では，状態 s が変動することはないため，状態 s については考えない．行動評価手法は加重平均手法である．ただし，自己の選択した学習法に対する評価式と他者情報に対する評価式が異なる (式(5.2), 式(5.3)). 式で異なっているのは，ステップサイズ・パラメータである．α の大きさで，自身の学習を重視するかどうか，γ の大きさで他者の情報を重視するかどうかが決まる.

自己の選択した学習法に対する評価式

$$Q_{n+1}^{mth}(m) = Q_n^{mth}(m) + \alpha\left(r_{n+1} - Q_n^{mth}(m)\right) \tag{5.2}$$

m ：自己が選択した学習法

r ：自己が得た報酬

α ：ステップサイズ・パラメータ（$0 \le \alpha \le 1$）

他者情報に対する評価式

$$Q_{t+1}^{mth}(m') = Q_t^{mth}(m') + \gamma\left(r'_{t+1} - Q_t^{mth}(m')\right) \tag{5.3}$$

m' ：他者が選択した学習法

r' ：他者が得た報酬

γ ：ステップサイズ・パラメータ（$0 \le \gamma \le 1$）

・学習パラメータに関する設定

（学習法実行部で使う各手法）

学習法実行部で使う各手法で設定する学習パラメータを表 5.6 に示す.

表 5.6　各手法の学習パラメータ

パラメータ名	数値
softmax 法　τ	0.1
ε - greedy 法　ε	0.1
追跡手法　β	0.1
強化比較　α	0.1
強化比較　β	0.1
強化比較　リファレンス報酬 $\bar{r}$	1
加重平均手法　α	0.08
Q 学習法　　α	0.05
Q 学習法　　β	0.01

（学習法決定部で使う学習パラメータ）

学習法決定部で使う学習パラメータを表 5.7 に示す.

表 5.7　学習法決定部の学習パラメータ

パラメータ名	数値
ε - greedy 法　ε	0.1
加重平均手法 : 自己の選択した学習法に対する評価式　α	0.1
加重平均手法 : 他者情報に対する評価式　γ	0.01

5.4　実験結果

5.4.1　学習法選択の推移について

実験結果を図 5.6～5.8 に示す．これらの図は Th，Amp 毎に最も選択されている手法を色で表したものである．図の右側にあるカラーバーとカラーバー横の数字は学習手法を表している．カラーバー横の数字と各学習手法の対応表を表 5.8 に示す．

学習が進む毎に手法の分布が図 5.6(a),(b),(c)，図 5.7(a)より，学習初期は，エージェントはまだ学習が浅いため，選択する手法は適当なものを選んでいる．図 5.7(b)，(c)のように学習が進んでくると，少しずつ各 Th，Amp で選択される手法がある程度決まってくる．左下から左上，左下から右下の領域については softmax 法が選択されるようになっている．また，左上から右下にかけてはε - greedy 法と追跡手法が混在している．右上の領域は広範囲にε - greedy 法が選択される．

学習が終了した試行 30000 回の結果である図 5.8(c)の図から学習法の選択領域がどのようになったかを図 5.9 を用いて説明する．

・図 5.9 の桃色で囲まれた領域について

この領域では主に softmax 法が選択されている．印刷の都合上，濃淡がはっきりしないためわかりにくいが，濃い紫色（■）は softmax 法＋加重平均手法，薄い紫色（■）は softmax 法＋Q 学習法である．softmax 法＋標本平均手法（■）はあまり選択されていない．

・図 5.9 の水色で囲まれた領域について

この領域では，ε - greedy 法と追跡手法が混在している．赤系の色の濃淡がはっきりしないが，主に赤紫色（■）の箇所はε - greedy 法＋加重平均手法，赤色の箇所（■）はε - greedy 法＋Q 学習法が選択されている．濃い黄色（■）の箇所は追跡手法＋加重平均手法，オレンジの箇所（■）は追跡手法＋標本平均手法が選択されている．

・図 5.9 の黄緑色で囲まれた領域について

この領域ではε - greedy 法が選択されている．薄紫（■）の箇所はε - greedy 法＋標本平均手法，赤紫色の箇所（■）はε - greedy 法＋加重平均手法，赤色の箇所（■）はε - geedy 法＋Q 学習法が選択されている．

5.4.2　獲得報酬量について

平均獲得報酬量について，次の２つの場合について比較を行った．比較は試行 29000 回～30000 回の 1000 回で得られた平均獲得報酬の差とする．

１．本システムとコミュニケーションなし（自身の経験のみで学習する）の場合について比較する．

２．本システムと各学習法固定の場合について比較する．

１．本システムとコミュニケーションなしの場合

比較結果を図 5.10 に示す．コミュニケーションなしの場合とは，全体的に 0.03 ポイント程度プラスの結果となった．

２．本システムと各学習法固定の場合

各学習手法との比較結果を図 5.11～5.20 に示す．なお，softmax 法＋加重平均手法など一部の学習法では差がわかりにくいため, Z 軸の範囲を変更したものを用意した．

・softmax 法＋標本平均手法（図 5.11）

0.12～0.18 ポイントプラスの結果となっている．Th，Amp の大きい領域に行くほど差が小さくなっている．

・softmax 法＋加重平均手法（図 5.12）

Amp が小さい領域(図中の紫がかった領域)は平均獲得報酬の差がほとんどない．赤色の領域は 0.01 ポイント, オレンジの領域は 0.02 ポイント程度プラスの結果となっている．

・softmax 法＋Q 学習法（図 5.13）

Amp が小さい領域（図中の紫がかった領域）は平均獲得報酬の差がほとんどないか 0.005 ポイント程度マイナスの結果となっている．赤色の領域は 0.01 ポイント，オレンジの領域は 0.02 ポイント程度プラスの結果となっている．

・ε - greedy 法＋標本平均手法（図 5.14）

Th または Amp が小さい領域については 0.17 ポイント程度プラスの結果となっている. Th と Amp の両方が 1 に向かうほど平均獲得報酬の差は小さくなり, Th, Amp が 1 付近では 0.12 ポイント程度プラスの結果となっている．

・ε - greedy 法＋加重平均手法（図 5.15）

Th，Amp の小さい領域（図中の紫がかった領域）は平均獲得報酬の差がほとんどない，もしくは 0.001 ポイント程度マイナスになっている．Amp が 0.2 以下の箇所（図中の濃い赤色の領域)は 0.005～0.01 ポイントプラスの結果となっている. また，オレンジの領域は 0.02 ポイント，濃い黄色の領域は 0.025 ポイントプラスの結果となっている．

・ε - greedy 法＋Q 学習法（図 5.16）

Th，Amp の小さい領域（図中の紫がかった領域）は平均獲得報酬の差がほとんどない，もしくは 0.001 ポイント程度マイナスになっている．Amp が 0.2 以下の箇所（図中の濃い赤色の領域)は 0.005～0.01 ポイントプラスの結果となっている. また，オレンジの領域は 0.02 ポイント，濃い黄色の領域は 0.025 ポイントプラスの結果となっている.

・追跡手法＋標本平均手法（図 5.17）

Th または Amp が小さい領域については 0.17 ポイント程度プラスの結果となっている. Th と Amp の両方が 1 に向かうほど平均獲得報酬の差は小さくなり, Th, Amp が 1 付近では 0.12 ポイント程度プラスの結果となっている.

・追跡手法＋加重平均手法（図 5.18）

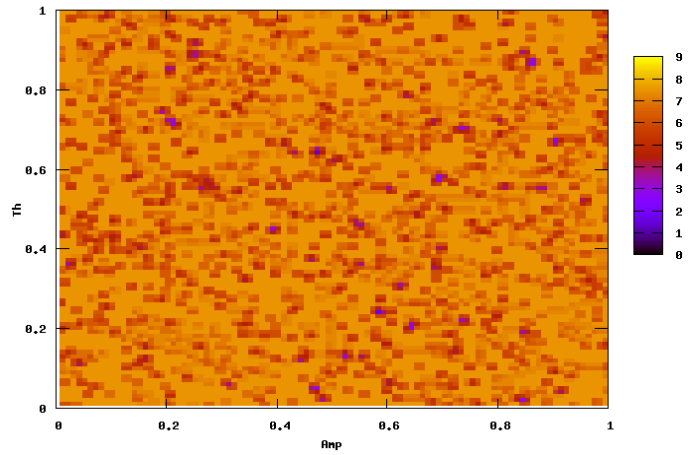
Th，Amp の小さい領域（図中の濃いオレンジの領域）は 0.08 ポイント程度プラスの結果となっている．また，Th，Amp がそれぞれ 1 に近づくほど平均獲得報酬の差が大きくなり，Th，Amp が 1 付近では 0.12 ポイント程度プラスの結果になっている.

・追跡手法＋Q 学習法（図 5.19）

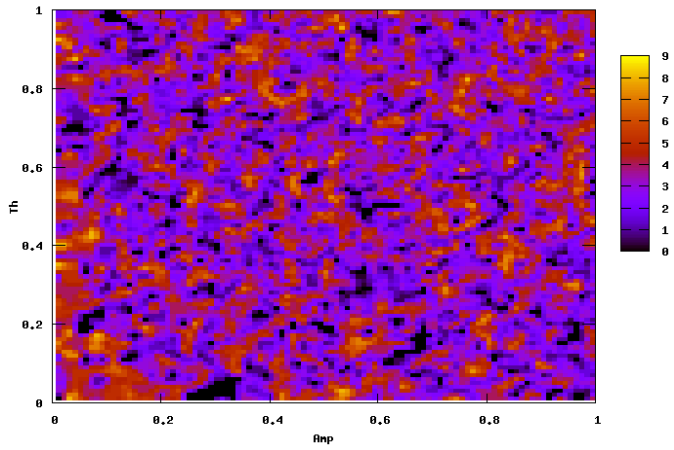
Th，Amp の小さい領域（図中の濃いオレンジの領域）は 0.08 ポイント程度プラスの結果となっている．また，Th，Amp がそれぞれ 1 に近づくほど平均獲得報酬の差が大きくなり，Th，Amp が 1 付近では 0.12 ポイント程度プラスの結果になっている.

・強化比較法（図 5.20）

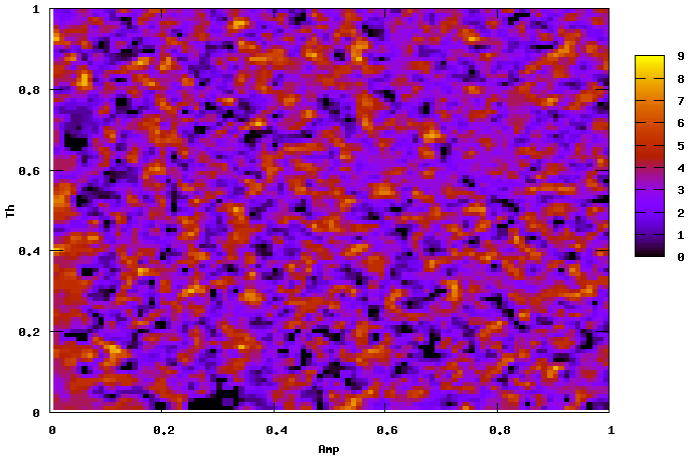
Th，Amp の小さい領域（図中のオレンジの領域）は 0.18 ポイント程度プラスの結果となっている．また，Th，Amp がそれぞれ 1 に近づくほど平均獲得報酬の差が小さくなり, Th, Amp が 1 付近では 0.11 ポイント程度プラスの結果になっている.



（a）試行 0 回目

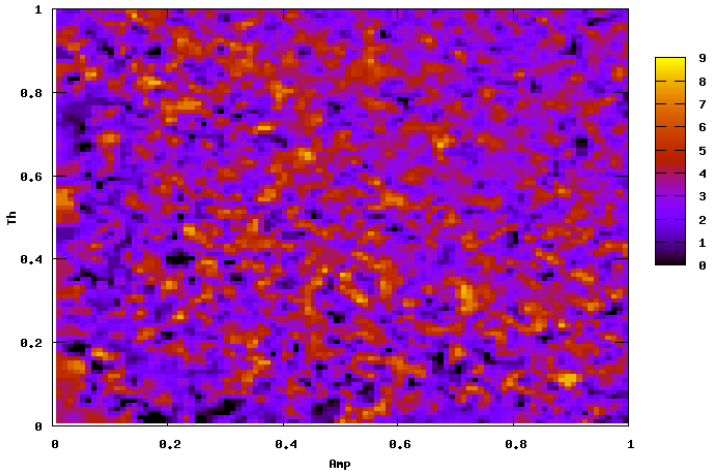


（b）試行 500 回目

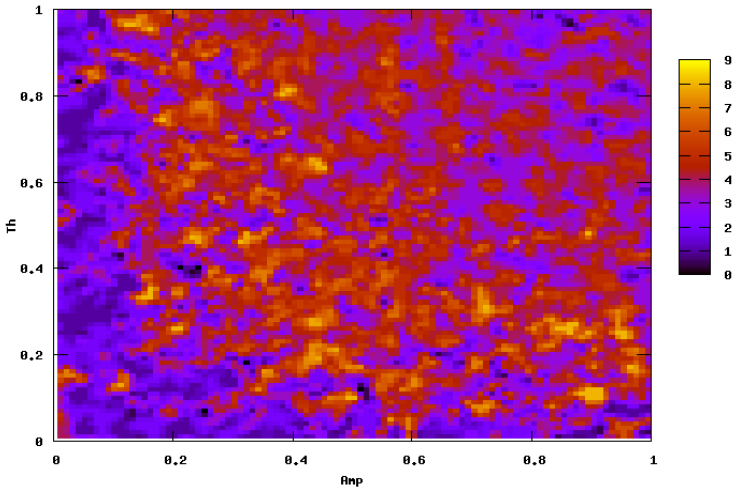


（c）試行 1000 回目

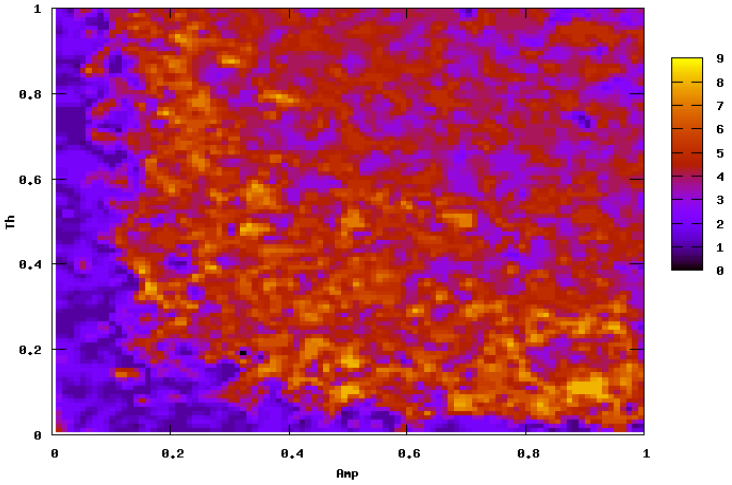
図 5.6　試行 0～1000 回までの選択手法の推移



（a）試行 2000 回目

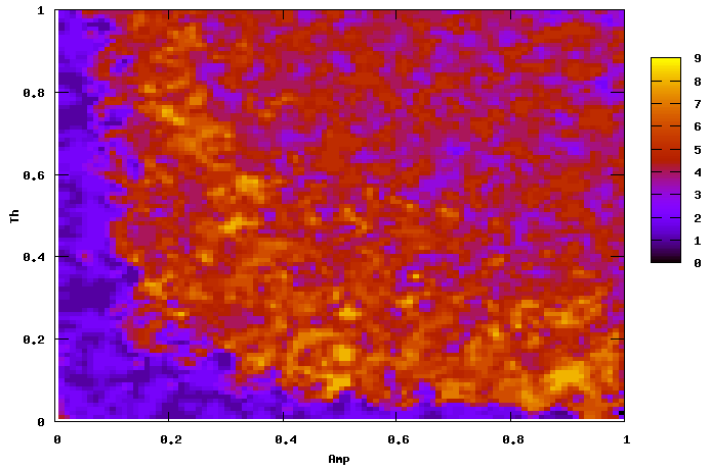


（b）試行 5000 回目

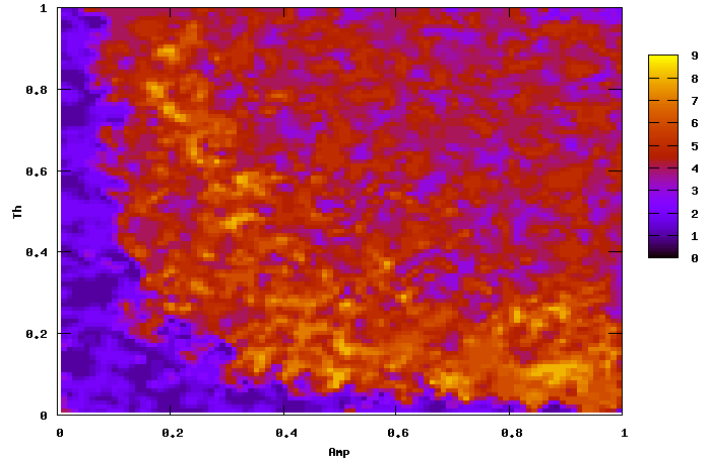


（c）試行 10000 回目

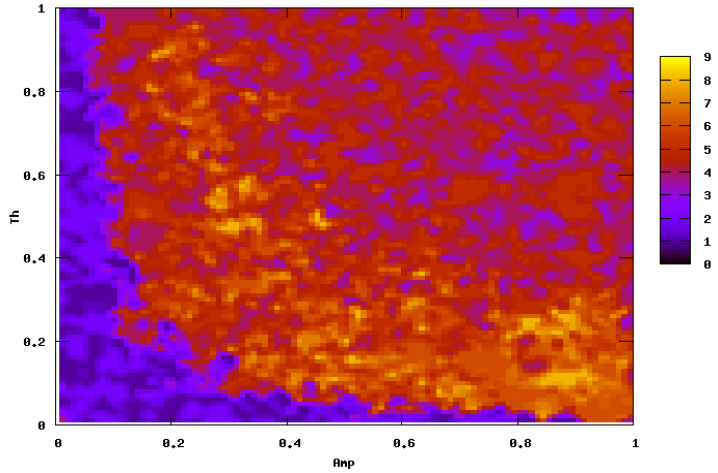
図 5.7　試行 2000～10000 回目の選択手法の推移



（a）試行 15000 回目



（b）試行 20000 回目



（c）試行 30000 回目

図 5.8　試行 15000～30000 回目の選択手法の推移

表 5.8　カラーバー横の数字と学習法の対応表

学習法	カラーバーの数字
softmax 法＋標本平均手法	0
softmax 法＋加重平均手法	1
softmax 法＋Q 学習法	2
ε - greedy 法＋標本平均手法	3
ε - greedy 法＋加重平均手法	4
ε - greedy 法＋Q 学習法	5
追跡手法＋標本平均手法	6
追跡手法＋加重平均手法	7
追跡手法＋Q 学習法	8
強化比較法	9

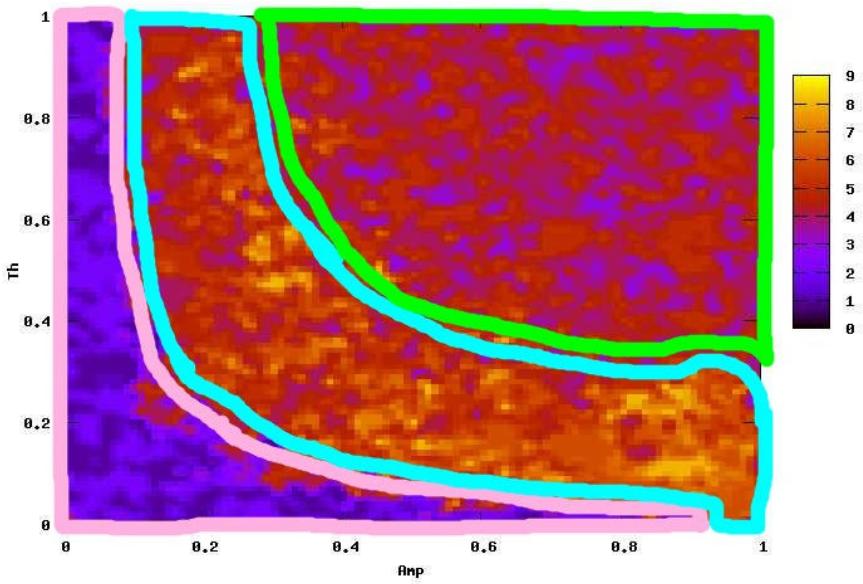


図 5.9　学習法選択領域

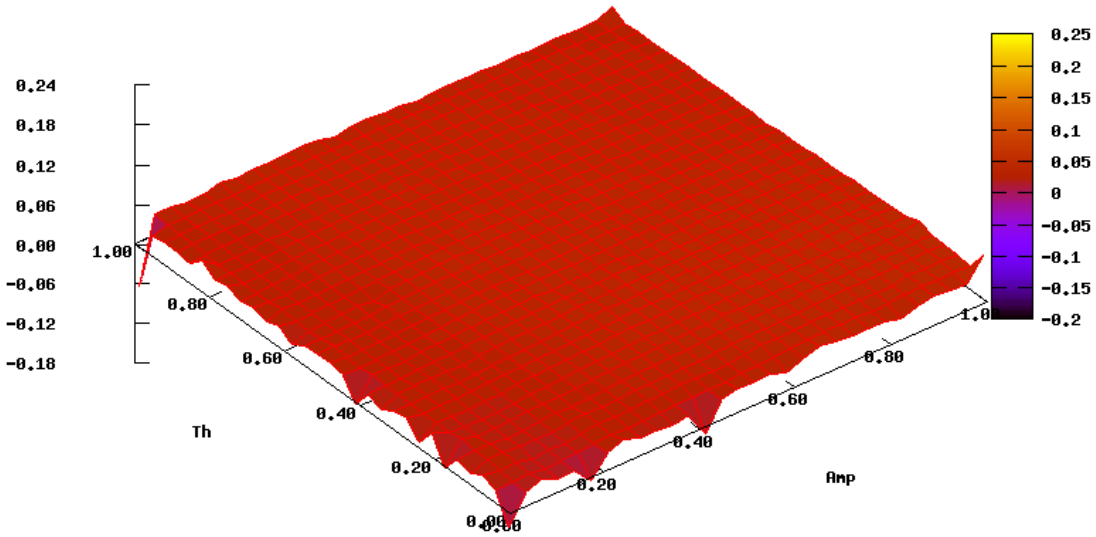


図 5.10　提案システムとコミュニケーションなしの場合との平均獲得報酬の差

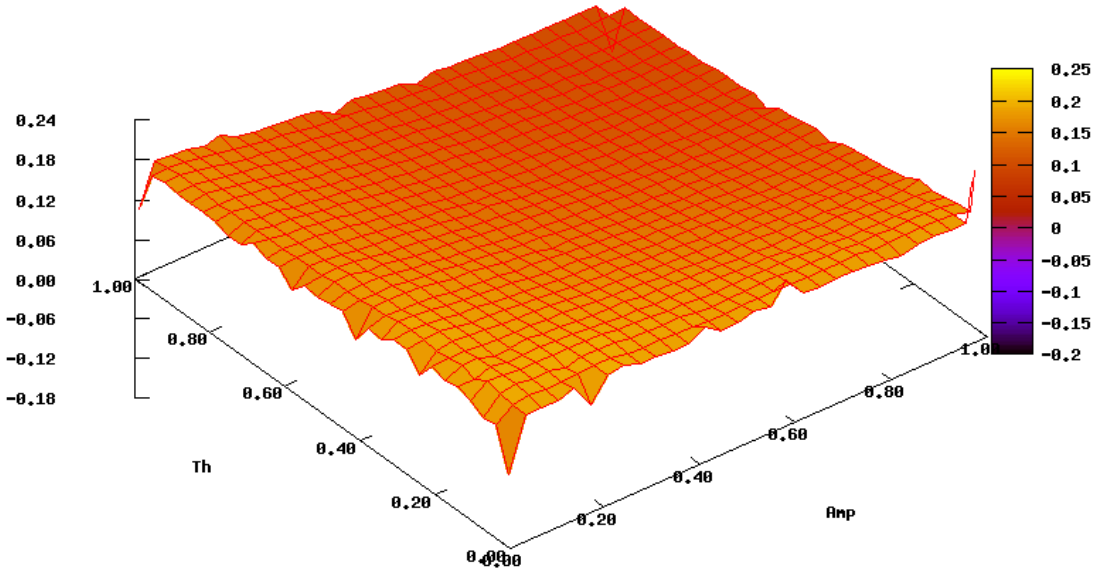
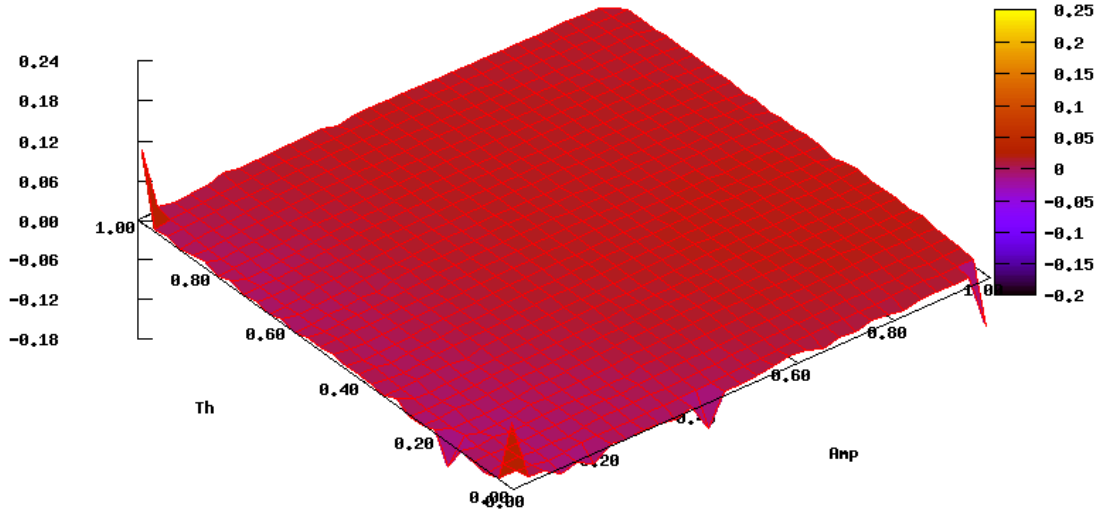
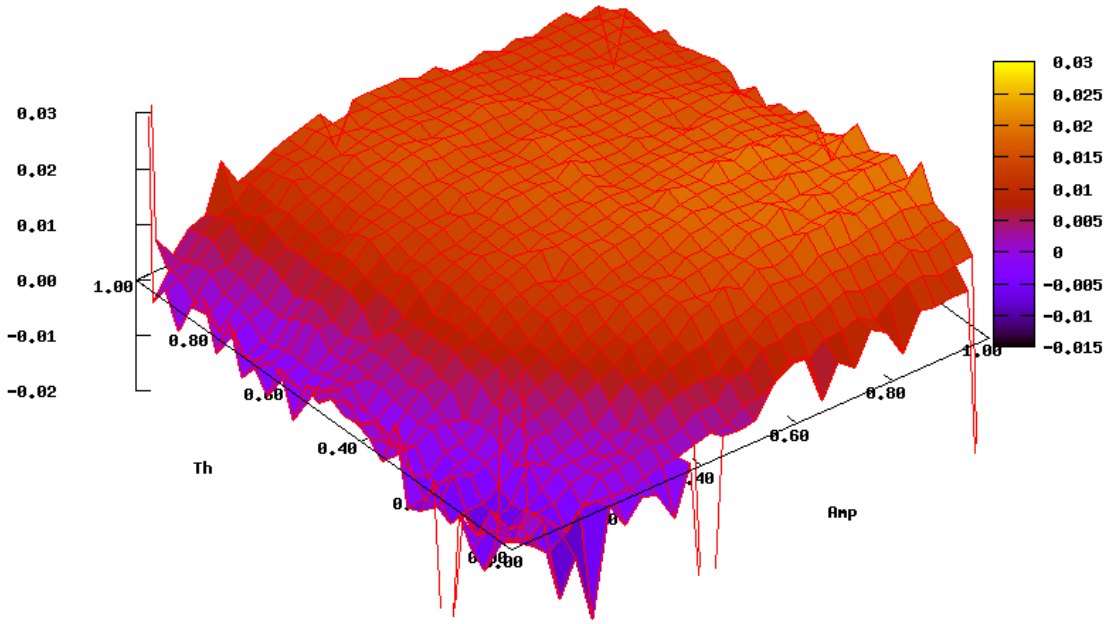


図 5.11　提案システムと softmax 法＋標本平均手法との平均獲得報酬の差

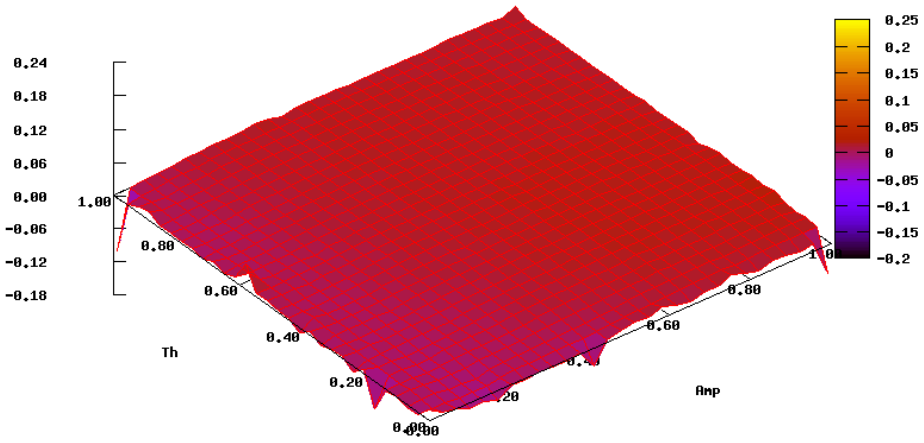


（a）Z 軸範囲-0.18～0.24

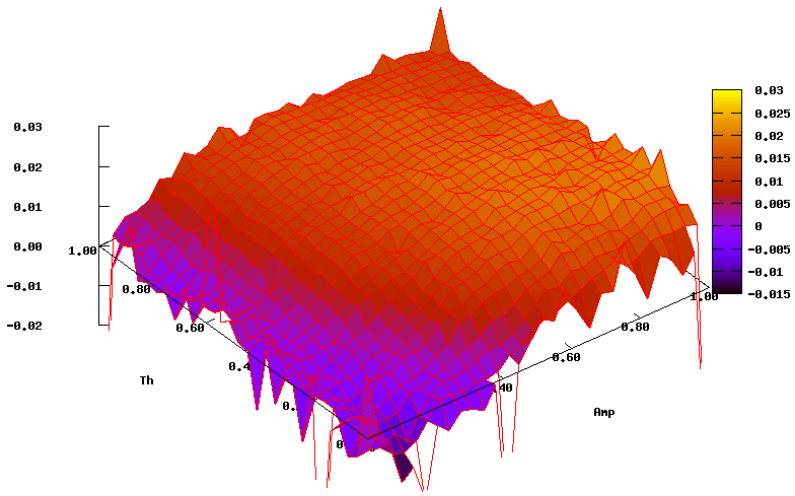


（b）Z 軸範囲-0.02～0.03

図 5.12　提案システムと softmax 法＋加重平均手法との平均獲得報酬の差



（a）Z 軸範囲-0.18～0.24



（b）Z 軸範囲-0.02～0.03

図 5.13　提案システムと softmax 法＋Q 学習法との平均獲得報酬の差

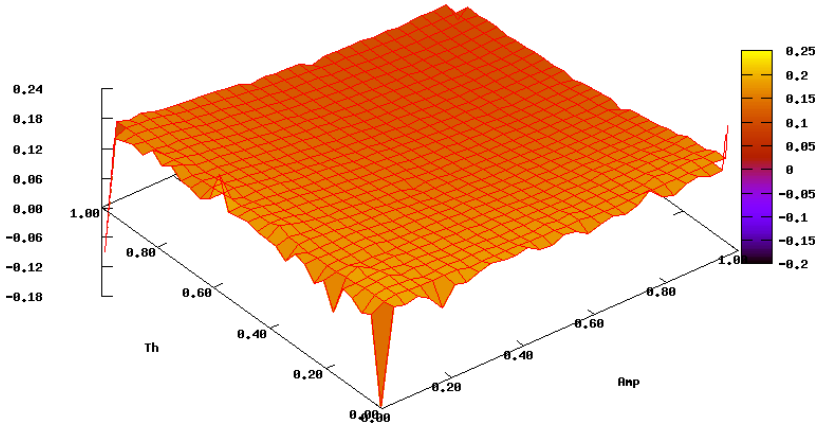
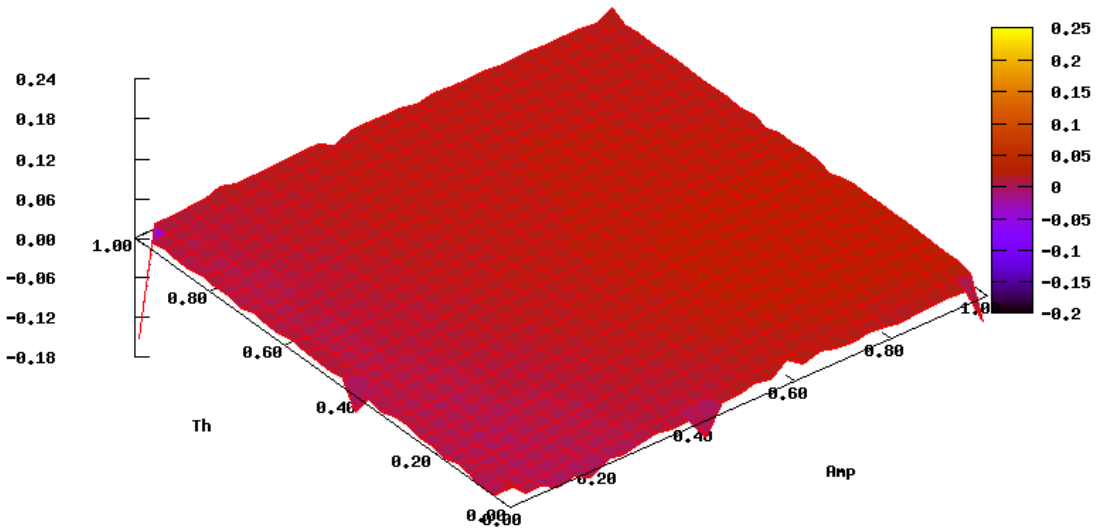
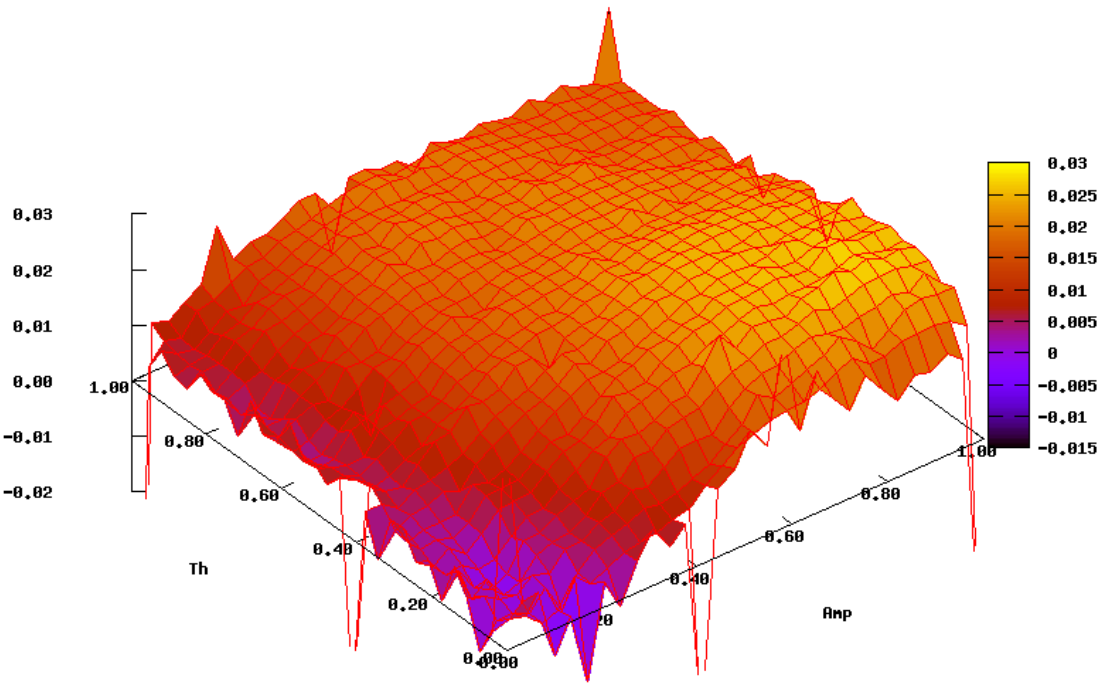


図 5.14　提案システムと ε - greedy 法＋標本平均手法との平均獲得報酬の差

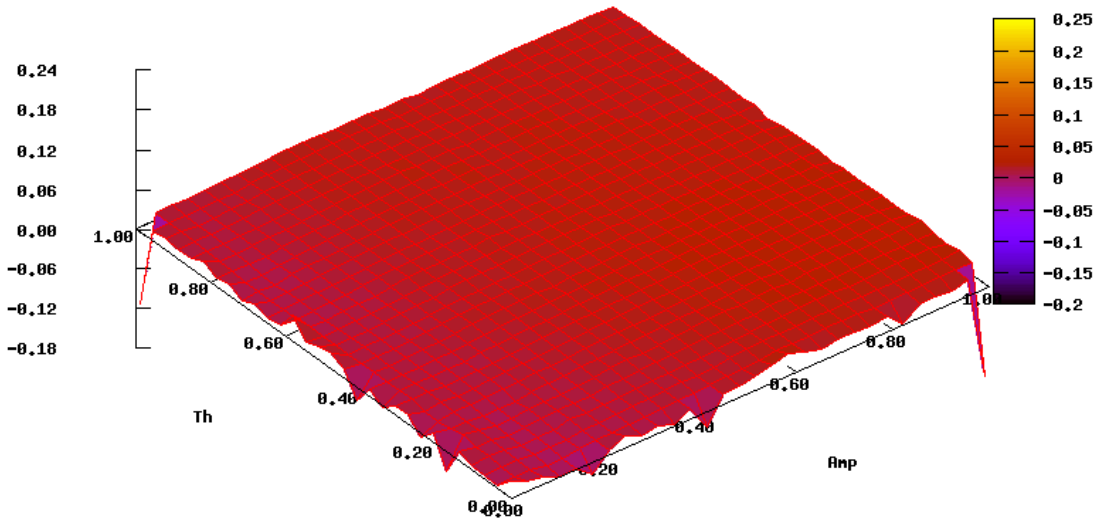


（a）Z 軸範囲-0.18～0.24

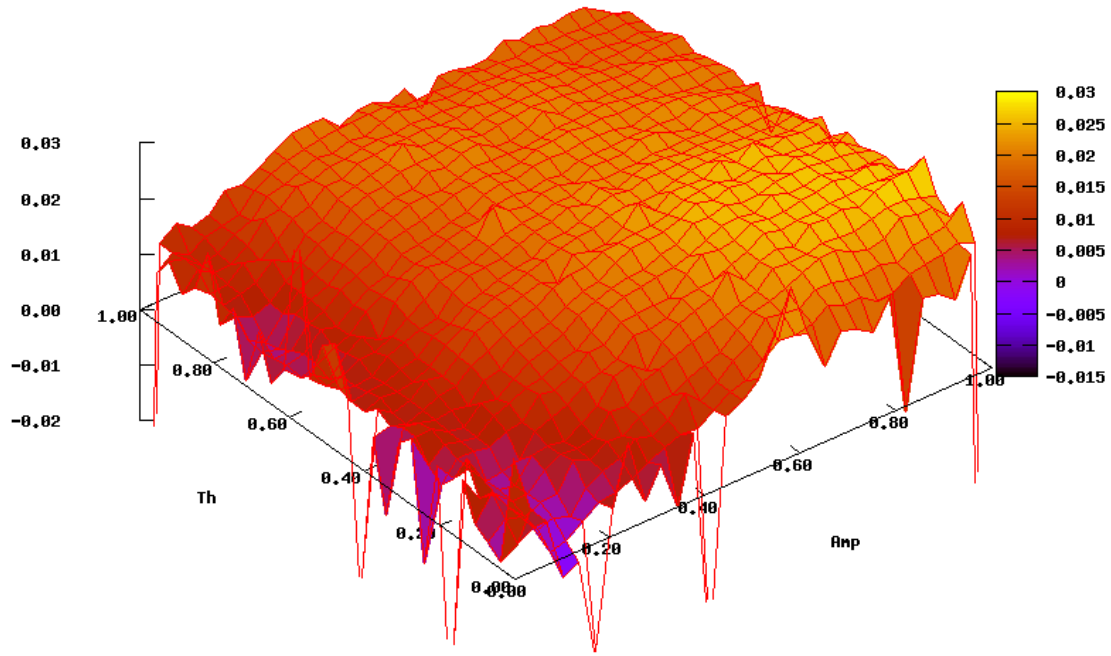


（b）Z 軸範囲-0.02～0.03

図 5.15　提案システムと ε - greedy 法＋加重平均手法との平均獲得報酬の差



（a）Z 軸範囲-0.18～0.24



（b）Z 軸範囲-0.02～0.03

図 5.16　提案システムと ε - greedy＋Q 学習法との平均獲得報酬の差

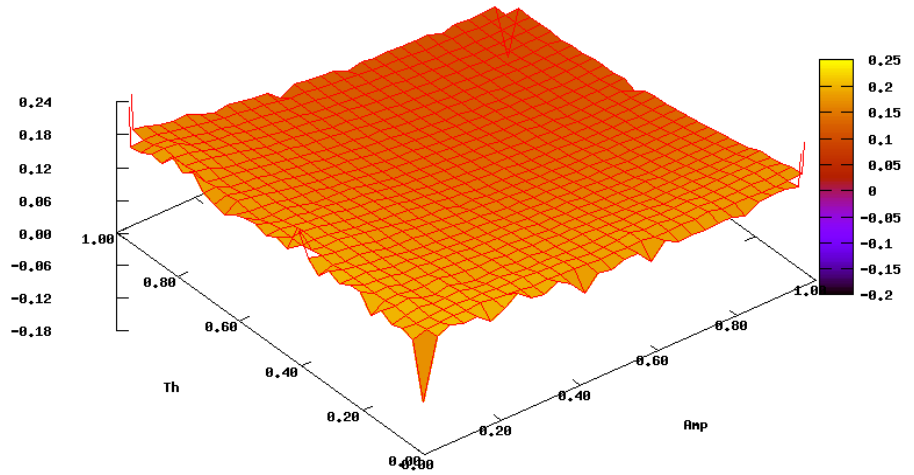


図 5.17　提案システムと追跡手法＋標本平均手法との平均獲得報酬の差

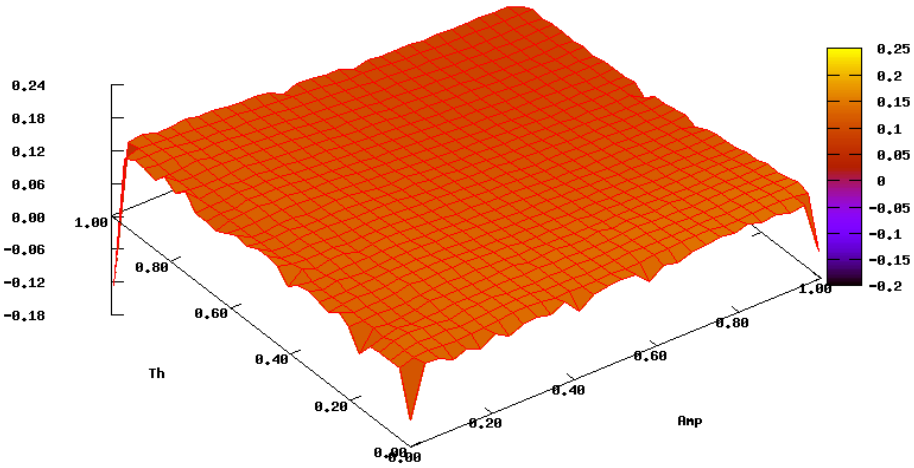


図 5.18　提案システムと追跡手法＋加重平均手法との平均獲得報酬の差

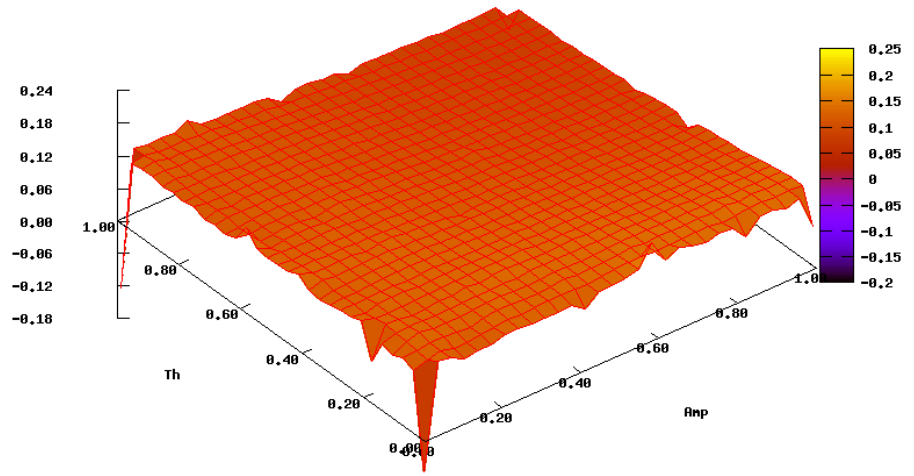


図 5.19　提案システムと追跡手法＋Q 学習法との平均獲得報酬の差

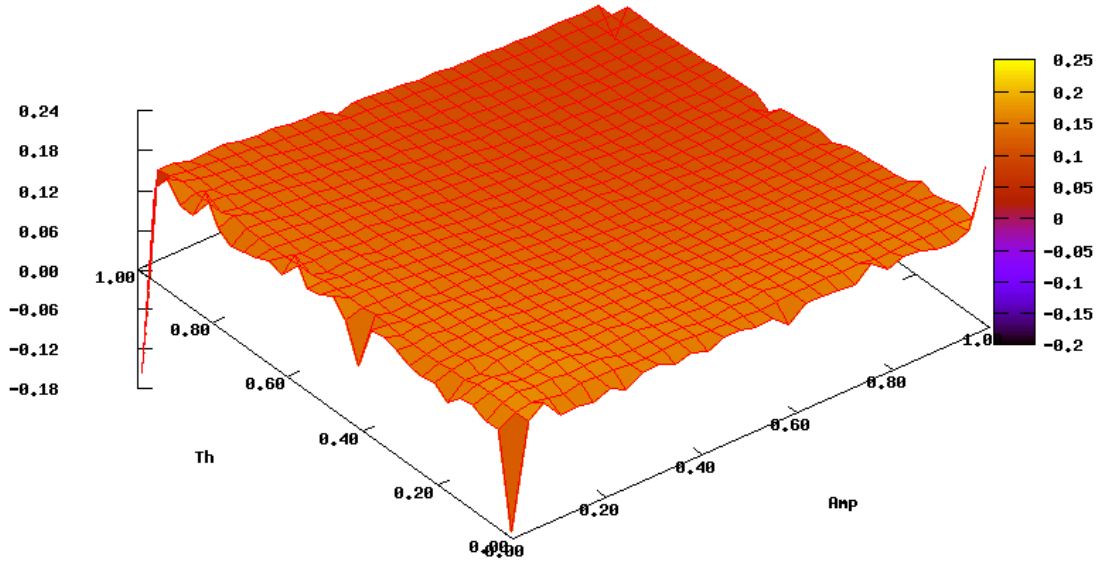


図 5.20　提案システムと強化比較法との平均獲得報酬の差

5.5　実験考察

5.5.1　学習法選択について

学習法選択については，図 5.8 より，試行数が 15000 回以降選択手法の分布にほとんど変化が見られなくなったため，試行 30000 回目では学習が安定したといえる．そのため，試行 30000 回目での学習法の分布について考察する．

- **図 5.9 の全ての領域に言えること**

  図 5.9 の桃色，水色，黄緑色の領域の分かれ方は，行動選択手法によって決まる．また，それぞれの領域内では，行動評価手法によって分かれている．

  行動評価手法の面では，全体的に標本平均手法の選択があまりされていない．標本平均手法は単純平均によって評価値を求める．そのため，学習が進むほど標本数が増え，新しく獲得した報酬が反映されにくくなる．学習が進んだ状態で獲得した報酬ほど重みを持たせ，評価した方がよい場合が多い．したがって，重みをつける行動評価手法である加重平均手法と Q 学習法が標本平均手法より選択されることが多い．

- **図 5.9 の桃色で囲まれた領域について**

  この領域は，腕の当選確率の変動が起こりやすく，変動量が小さい領域

（$0.0 \leq Amp \leq 0.1$，$0.4 \leq Th \leq 1.0$），変動が起こりにくく，変動量も小さい領域（$0.0 \leq Amp \leq 0.39$，$0.0 \leq Th \leq 0.2$），変動が起こりにくく，変動量が大きい（$0.4 \leq Amp \leq 0.8$，$0.0 \leq Th \leq 0.1$）．このような領域は，変動が多くても，変動量が少なく初期確率から変化しにくい．また，変動量が多くても変動が起こりにくい．そのため，この領域では確定的な手法が適している．本実験の softmax 法のパラメータ τ は 0.1 なので，softmax 法は確定的手法である greedy 法に近い挙動をする．そのため，この領域では softmax 法が選択されている．

- **図 5.9 の水色で囲まれた領域について**

  この領域は，腕の当選確率の変動が多く，変動量があまり多くない領域（$0.2 \leq Amp \leq 0.4$，$0.4 \leq Th \leq 1.0$）と，変動があまり多くなく，変動量が大きい領域（$0.41 \leq Amp \leq 1.0$，$0.0 \leq Th \leq 0.39$）で構成されている．このような領域では，腕の当選確率の変化の仕方が若干緩やかである．そのため，ある程度確率的な手法が適していると考える．この領域では，追跡手法と ε - greedy 法が選ばれている．追跡手法はある程度確率的に動くが，環境変化への追従性はあまり高くない．そのため，変動頻度の大きい領域（$0.2 \leq Amp \leq 0.4$，$0.4 \leq Th \leq 1.0$）ではあまり選択されていないが，変動頻度の小さい領域（$0.41 \leq Amp \leq 1.0$，$0.0 \leq Th \leq 0.39$）で選択されることが多くなっている．また，本実験では，行動学習部の ε - greedy 法の ε の値は 0.1 なので，10 回試行に 1 回はランダムに腕を選択する．このため，追跡手法よりさまざまな腕を試すことが可能であったため，変動頻度の大きい領域である程度選択されていると考える．

- **図 5.9 の黄緑色で囲まれた領域について**

  この領域では，腕の当選確率が変動しやすく，変動量も多い．よって，当選確率の高い腕は頻繁に変わる．このような領域では，greedy に行動しつつもある程度ランダムな挙動をする手法が適していると考える．これは，頻繁に当選確率の高い腕が変わるため，現在選択している腕の当選確率が下がっても，またすぐに腕の当選確率が高くなる場合多いためである．また，頻繁に当選確率の高い腕が変わることから，ランダムな挙動をした方がよい場合がある．

  本実験では，行動学習部の ε - greedy 法の ε の値は 0.1 なので，10 回試行に 1 回はランダムに腕を選択する．これは，ある程度ランダムな挙動をする手法といえる．したがって，この領域では ε - greedy 法が選択されていると考える．

以上より，各エージェントは自身が直面する環境に合った学習法を選択していると考える．

5.5.2　平均獲得報酬量について

図 5.8 より，試行数が 15000 回以降選択手法の分布にほとんど変化が見られなくなったため，試行 30000 回目では学習が安定したといえる．そのため，試行 29000～30000 回の 1000 回の獲得平均報酬の比較は妥当だと考える．よってここからは，獲得平均報酬から提案システムの有効性を考察する．

図5.10からコミュニケーションなしの場合と比較して全体的に0.03ポイント程度提案システムの方がプラスの結果であったことから，単体での学習よりも効果的に学習できているといえる．

学習法固定の場合については，図 5.11 の softmax 法＋標本平均手法，図 5.14 のε - greedy 法＋標本平均手法, 図 5.17～図 5.20 の追跡手法を使った学習法と強化比較法に関しては，全体的にプラスの結果となっている．

図 5.14(b)，図 5.15(b)のε - greedy 法＋加重平均手法，ε - greedy 法＋Q 学習法では，Th，Amp が共に低い箇所を中心に 0.001 ポイント程度マイナスの結果であった．また，図 5.12(b)，図 5.13(b)の softmax 法＋加重平均手法，softmax 法＋Q 学習法については同様に 0.005 ポイント程度マイナスの結果になっている．これは，学習法学習部の選択手法にε - greedy 法を採用しているためと考えられる．ε - greedy 法は確率εでランダムに他の学習法を選択する．そのため，価値の低い学習法を選択してしまい，報酬が得られないということが起こる．その結果，得られる報酬が若干少なくなり，マイナスの結果になったと考えられる．実際，図 5.11 のように Th，Amp が共に低い場所は，初期確率からの変動が少ないため， softmax 法＋加重平均や softmax 法 Q 学習法が多く選択されているが，他の手法も選択されている．しかし，結果がマイナスなのは Th，Amp が共に小さい領域のみであり，他の領域は軒並みプラスの結果となっている．また，実環境では，定常環境であることは少ないため，より効率よく学習するためには，自身が直面する環境に合った学習法を学習するべきであると考える．よって，手法固定の場合と比べて多少マイナスがあっても，提案システムの方がよいと考える．以上から，提案システムは有効であると考える．

5.5　まとめ

第 5 章では，提案システムの有効性の検証として， 非定常環境での N 本腕バンディット問題を対象とした実験を行った．実験環境として，変動頻度と変動振幅からなる環境マップを生成し，その環境マップに基づいてバンディットの腕の当選確率を変化するような設定をした．そのようなバンディットに対してエージェントは提案システムを用いて適した学習法の学習を行う．実験結果は，コミュニケーションなしの手法と学習法固定の場合との平均獲得報酬値の比較から，提案システムのほうが高い平均獲得報酬値を記録したことから，提案システムが有効であることが検証できた．

第６章　結論

6.1　まとめ

群の中の個体の知能の発達として，コミュニケーションを用いて個体知能の発達を促進させるシステムの構築を目的とした．本論文では「情報の形式が個体間で固定」，「情報の処理・利用の仕方」が個体間で共通という条件から，「個体の身体構造が同一」，「目的・タスクが同一」という群の個体に注目した．このような群個体でコミュニケーションすることを考えると，「個々の状況・身体に非依存な情報」がコミュニケーションに有効な情報であると考えた．そして，そのような情報の中からコミュニケーションに用いる情報を「学習法」に設定した．学習法をコミュニケーションすることで，それぞれの個体が直面する環境に合った学習法を学習するシステム概念を示した．

本稿では，この概念に強化学習を適用したシステムの構築を行った．この提案システムの有効性を検証するため非定常環境Ｎ本腕バンディット問題を対象とした実験を行った．まず，学習法の選択の推移を観察すると，試行 15000 回以降は選択している手法の分布が安定したため，エージェントがそれぞれ直面する環境に合った学習法を学習していることが分かった．次に，コミュニケーションのなしの場合と人間が学習法を設定した場合の２つの場合と平均獲得報酬量で比較した結果，全体的に提案システムの方がプラスの結果が出た．これら２つのことより，提案したシステムの有効性を検証することが出来た．

6.2　今後の課題

今後の課題を以下に示す．

１．他のタスクにも提案システムを適用する．

今回の実験では，Ｎ本腕バンディット問題に対して提案システムの有効性を検証した．今後はＮ本腕バンディット以外のタスクに対しても提案システムの有効性を検証する．

２．学習法の学習で扱う学習法に他の学習法を適用する

今回は強化学習について適用したが，ニューラルネットワークや遺伝的アルゴリズムといった強化学習以外の学習法についても適用し，有効性を検証する．

３．学習法以外の情報についてもコミュニケーションする情報として扱う．

今回は学習法についてコミュニケーションし，個体の知能を発達させるシステムを提案したが，他の情報についてコミュニケーションするシステムを考える．

4．実ロボットへの適用

コンピュータシミュレーションだけでなく，実ロボットを用いて実環境でコミュニケーションを行い，個体の知能を発達させる実験を行いたい．

5．より高度なコミュニケーションの考察

- **他タスク・他目的の個体同士でのコミュニケーション**

今回は出来るだけ簡単なコミュニケーションを考えた，同タスク・同目的のもと，情報の処理の仕方も各個体で同一のものとした．しかし，より高度なコミュニケーションでは，異なったタスク・目的でもコミュニケーションを行っている．この場合のコミュニケーションは，得られた情報の処理の仕方も自身で考えて行うことが必要である．これはタスク・目的が異なると，同じ情報でも個体によって評価が変わってしまうためである．例えば，ゴールが複数ある迷路タスクでゴールにいたる最短経路に関する情報を個体間でコミュニケーションすることを考える（図 6.1）．

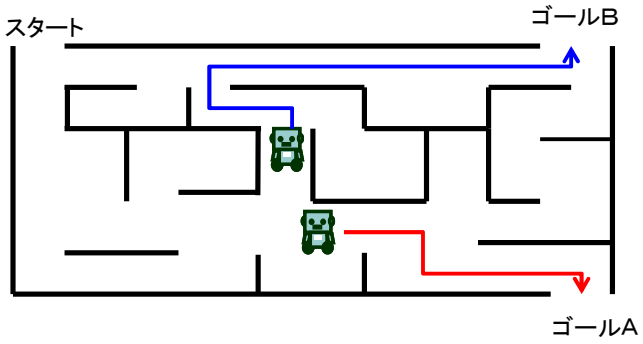


図 6.1　目的が異なる場合

この場合，個体ごとに目標となるゴールが異なるため，その情報は自身にとって悪影響を与え，その情報に対する評価を下げることになる．しかし．このような情報は自身の情報の処理の仕方によって，評価を下げることなく利用することが出来る．

例えば，相手が自身と同一のゴールの場合，情報は自身の最短経路に関する情報なので，自身が得た情報と同じ評価の仕方をする．しかし，自身と異なったゴールの場合，他者情報は自身にとっては最短経路ではない．そのような場合，情報の評価の仕方は最短経路ではない情報として評価する必要がある．つまり，「最短経路ではないため利用しない」ように学習する．そうすれば，最短経路でない情報も利用することが可能になる．このように他タスク・他目的の情報を利用するためには，自身が他者のタスク・目的を知っている必要がある．その上で，他者情報の処理方法を考える．

- **情報の取捨選択について**

人間は，受け取った情報を取捨選択し，自身に必要な情報を抽出し利用している．ロボットもこのような人間の情報の扱い方を実装することでより柔軟なコミュニケーションが可能となる．

ロボットの情報は，決められた形式で記録されている．情報の取捨選択は，この決められた形式で記録された情報から，必要な情報を抽出するということである．例を図 6.2 に示す．このような形式で記述されたデータから，必要なデータ（青枠で囲った箇所）のみを抽出し，利用する．そうすることで，自身の直面する状況に応じて必要なデータを利用することが出来る．

時間	温度	距離	音
1	38.2	236	56
2	38.5	436	70
3	40.1	668	86
4	41.0	710	89

図 6.2　データの抽出

謝辞

本論文を結ぶにあたり，日頃より懇切なるご指導を賜りました倉重健太郎先生に深く感謝の意を表します．また，ご指導，ご助言を頂いた畑中雅彦先生，本田泰先生，須藤秀紹先生，渡部修先生に感謝の意を表します．そして，論文の査読や助言をして頂いた院生の尾上由希子さん，プログラムや発表スライドで助言を頂いた同輩の池田憲弘君，幾世橋将文君に感謝いたします．

参考文献

[1] 小高 知宏著：はじめての AI プログラミング C 言語で作る人工知能と人工無能，6.3 ニューラルネットワーク，オーム社（2006）

[2] R.S. Sutton and A.G. Barto. “Reinforcement Learning: An Introduction.” MIT Press,（1998）.（邦訳：“強化学習”，三上，皆川 訳，森北出版，（2001））

[3] 田島 克樹 : 複数台移動ロボットを用いた追い込み行動の研究―簡易機能による追い込み行動の実現，東京電気大学大学院工学研究科電子工学専攻修士論文（2005）

[4] 矢萩 孝志 : 知能表現を行う群ロボットを目指した小型ロボットの製作，高知工科大学知能機械システム工学科学士論文（2003）

[5] 松本 浩平 : 複数台移動ロボットによる協調に関する基礎研究―小型ロボット MK-01X の基本走行の実現―，東京電機大学大学院工学研究科電子工学専攻修士論文（2005）

[6] 松山 隆司，浮田 宗伯：能動視覚エージェント群による協調追跡，日本ロボット学会誌, Vol.19, No.4, pp.25-31, 2001

[7] D. Cliff and G. F. Miller. Co-evolution of pursuit and evasion II：Simulation methods and results. In Proc.of the 4th International Conference on Simulation of Adaptive Behavior: From Animals to Animats 4., pages506–515, 1996.

図の参考

[1] 東京工業大学 大学院総合理工学研究科 知能システム科学専攻　創発システム講座 創発的機能形成分野　小林重信 研究室ホームページ　強化学習の基礎理論より
http://www.fe.dis.titech.ac.jp/research/rl/RL-Tut2.html