目次

第 1 章　序論

1.1　背景

利用され始めた当初のロボットは，工場の生産ラインなど，環境内の限られた場所で人間の代わりに単純な繰り返し作業を行っていた．このようなロボットは，設計者が事前に想定した各状態に対して, 設定された動作を行う単純なものであった. 近年ではロボット技術の進歩によって，ロボットはより複雑な作業を行うことが可能になった．それに伴い，家庭環境やオフィスなどの身近な環境や，災害現場などの人間が作業を行うことが困難な環境など，様々な環境において利用されるロボットの開発・研究が行われている．

このように近年，ロボットは様々な環境において利用されるようになり，利用され始めた当初のような，設計者が事前に設定した動作だけでは直面した状態に対応出来ない場面が出てきた．この理由として，近年のロボットが利用される環境は複雑であるため，設計者がロボットの直面する状態を全て想定することは困難であることが挙げられる．そこで，複雑な環境においても作業を行えるよう，機械学習によって自律的に環境に適応した動作を学習する方法が研究されている[1]-[3]．本論文では，実ロボットに用いられることも多い，強化学習[4][5]に注目する．強化学習は環境に対して試行錯誤を繰り返すことで，環境に適応した行動を学習する手法である．

1.2　ロボットの状態認識と学習の関係

複雑な環境において，ロボットがとった行動や他の要因，例えば環境内の人間の行動などの要因によって，ロボット自身やロボットの周囲の状態(以下，環境状態)は変化する．したがって，ロボットは複雑な環境下で作業を行うに当たり，環境状態の変化に適応した行動を学習する必要がある．ここで学習において，ロボットはセンサを用いて環境状態を観測し，得られたセンサ値から認識した状態(以下，観測状態)に対して適切な行動を学習していく(図 1.1)．したがって観測状態は，ロボットが適した行動を学習することが可能であるかを決める重要な要素であると考えられる．しかし複雑な環境において，センサを用いて環境状態を正確に認識することは困難である. 以降でその理由と, 複雑な環境におけるロボットの状態認識の問題について説明していく．

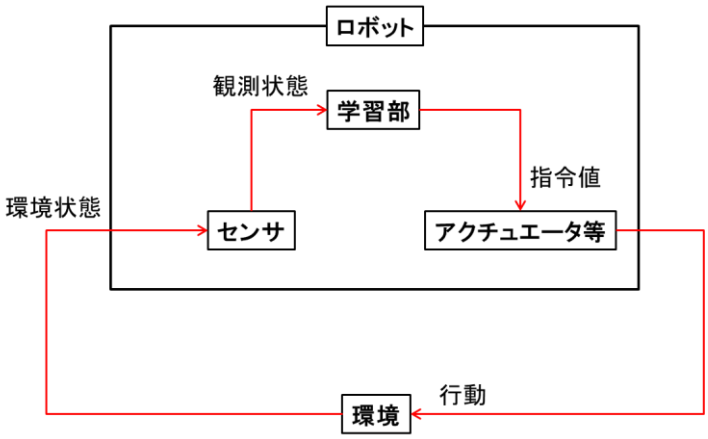

図 1.1　ロボットの認識と学習の関係

ロボットがセンサを用いて環境状態を認識するとき，その認識能力は搭載しているセンサによって決まる．一般にセンサの測定能力(測定出来る物理量，測定範囲など)には限界があり，その能力以上の物理量，例えば，測定出来ない物理量や測定範囲から外れた物理量などの測定は出来ない．つまりロボットがセンサを用いて正確に環境状態を認識するためには，測定する必要がある物理量やその範囲を設計者が事前に想定し，十分な能力，種類のセンサをロボットに搭載する必要がある．しかし複雑な環境において，ロボットの状態認識に必要な物理量や，その範囲を事前に想定することは困難である．このため複雑な環境において，ロボットに搭載されたセンサの測定能力は正確に環境状態を認識する上で不十分であり，観測状態と環境状態は一致しないことが多い．このとき，複数の環境状態において同じ観測状態が得られることが考えられる(図 1.2)．つまり，ある観測状態が複数の環境状態を混同してしまっている状況である．このような，不完全性が存在する状態認識は不完全知覚と呼ばれ，学習に及ぼす影響について様々な研究が行われている．逆に，状態認識に不完全性が存在しない場合，つまり，観測状態と環境状態が一致する場合を本論文では完全知覚とする．

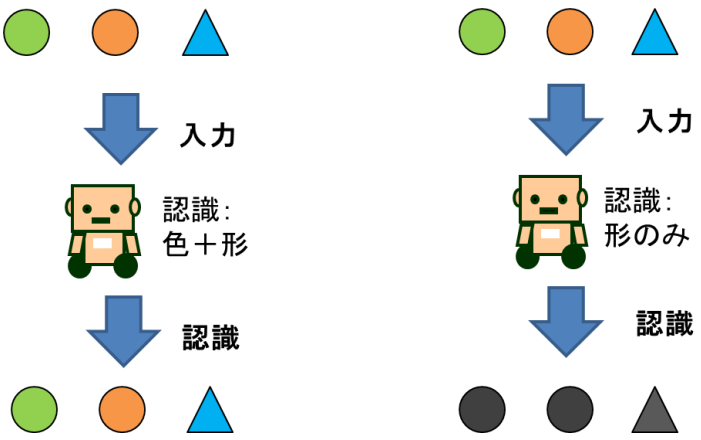

図 1.2　センサを用いた状態認識の不完全性

状態認識が不完全知覚である場合，複数の環境状態で同じ観測状態が得られる場合が考えられる．このとき，学習は環境状態ではなく観測状態に対して行われるため，不完全知覚が学習に悪影響を及ぼす可能性がある．例えば，障害物を認識して避ける動作を行うためには，距離センサを用いて障害物との距離を認識することが有効であると考えられる．しかし，距離センサの測定範囲が狭い場合には障害物がセンサの測定範囲に入らず，障害物の有無に関わらず「障害物はない」という観測状態しか得られないことが考えられる．この場合には障害物があることを認識することが出来ず，その結果，障害物を避ける動作を学習することも出来ないと考えられる（図 1.3）．

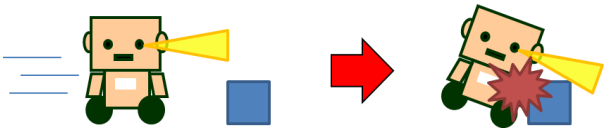

図 1.3　認識の不完全性と学習の関係

一方，状態認識が不完全知覚であっても，学習への悪影響がない場合もある．不完全知覚により，ある観測状態が複数の環境状態を混同していたとする．このとき，混同している全ての環境状態でとるべき行動が一致していれば，この観測状態に対してとるべき行動を学習出来る可能性は高い．したがって，このような場合には学習へ悪影響を及ぼす可能性は低いと考えられる．例えば迷路問題において，エージェントが周囲（上下左右のマス）の壁の有無をセンサにより観測し，状態認識を行うとする．このとき図 1.4 に示す環境において S はスタート位置，G はゴール位置を表している．また，矢印は適した行動を表しており，色のついた各マスは同じ色のマスで同じ観測状態が得られるマスである．このような設定において，複数の環境状態（この場合は迷路の各マス）で同じ観測状態を得ているため，環境状態と観測状態は一致していない．つまり，このエージェントの状態認識は不完全知覚である．さらに，矢印は各マスでゴールに到達するために移動するべき方向を表している．しかし，エージェントは複数のマスを混同している■と■では下，■と■では左へ移動することで，スタートからゴールに到達することが出来る．このように，状態認識が不完全知覚であっても学習に悪影響を及ぼさない場合もある．

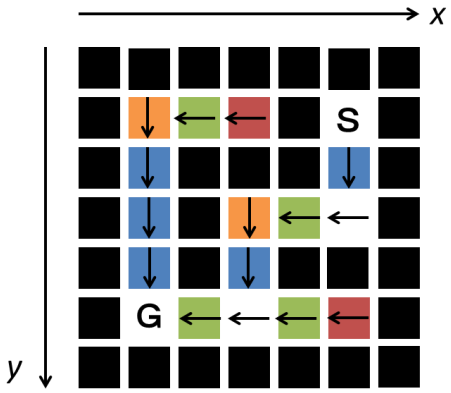

図 1.4　不完全知覚が学習に悪影響を及ぼさない環境の例

以上のことから本論文では，混同している複数の環境状態においてとるべき行動が複数ある場合に，不完全知覚が学習に悪影響を及ぼすと考える．このような状況を改善することで，学習における不完全知覚問題を解決することが出来ると考えられる．

1.3　先行研究とその問題点

不完全知覚を解決する方法については様々な研究がされている[5]-[9]．本論文では，その 1 つである「不完全知覚に対する状態認識法の提案–経験情報に基づく現状態の推定–」[10]を先行研究として扱う．先行研究はセンサを用いて状態を認識し，強化学習を行うロボットを対象としている．先行研究では，センサ情報のみを用いて状態認識を行っているために不完全知覚が起こると考え，センサ情報に加えてロボット自身の経験情報を状態認識に用いることで，不完全知覚の改善を図っている．先行研究の手法において，ロボットはセンサを用いて認識した状態が不完全知覚であるかを，自身の経験情報を用いて判断する．そして，不完全知覚であると判断したときにはその状態を，不完全知覚である状態と，経験情報と合わせた新たな状態の 2 つに細分化する．この新たな状態を知識化して記憶し，以降の状態認識に利用することで，不完全知覚で混同している状態を区別出来るようにしていく(図 1.5)．

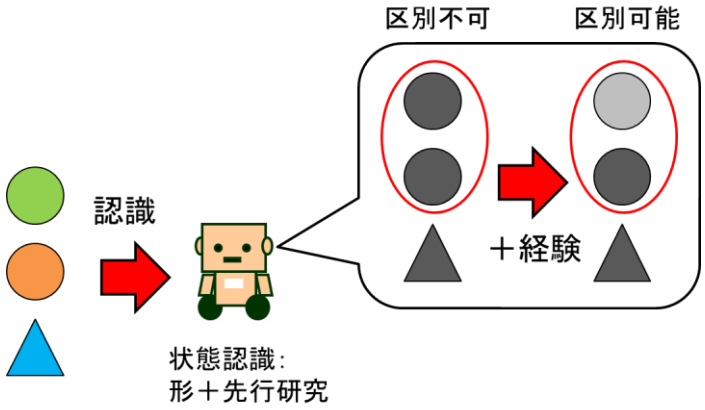

図 1.5　先行研究の手法を適用したロボットの状態認識

先行研究はシミュレーション実験により，この手法が不完全知覚を改善出来ることを示していた．その一方で，知識量の増加という問題が発生している．知識量の増加が問題になる理由として，ロボットのメモリは有限であるため，増加する知識を全て記憶することは出来ないという問題や，知識が増加することで認識出来る状態が増加し，学習時間が増加するという問題が挙げられる．本論文では知識量の増加という問題を解決するため，先行研究で行っている状態の細分化に注目する．

1.4　研究目的

本論文では先行研究と同様，学習に強化学習を用いる．その上で，先行研究の問題点を解決するため，先行研究で行っている状態の細分化に注目して改善を行う．それにより，知識量を抑制しつつ，強化学習における不完全知覚問題を解決する手法を提案することを目的とする．また，提案手法の有効性をシミュレーションによって示す．

1.5　アプローチ

先行研究には知識量の増加という問題が存在した．本論文では，知識量の増加を抑制するために先行研究の細分化に注目する．具体的には以下の 2 点に注目し，知識量を抑制する方法を考える．

・細分化を行う方法

・細分化を行う対象

まず，細分化を行う方法について説明する．先行研究の手法ではロボットが自身の経験情報を用いて，認識した状態が不完全知覚であるかを判断する．そして，不完全知覚であると判断した場合に状態を細分化していた．つまり，先行研究では認識した状態が不完全知覚かどうかで確定

的に細分化を行っていた．このため，先行研究の手法を適用したロボットは，不完全知覚である状態が多い複雑な環境において，頻繁に細分化を行い，知識が急激に増加すると考えられる．したがって本研究では，ロボットが自身の経験情報を用いて確率的に細分化を行うようにすることで，知識量の抑制を図る．

次に，細分化を行う対象について説明する．先行研究では，ロボットは自身の経験情報を用いて，認識した状態が不完全知覚であると判断した場合に細分化を行っていた．しかし 1.2 節で述べたように，認識した状態が不完全知覚であっても，混同している環境状態で適した行動が一致していれば学習に悪影響はないと考えられる．したがって，先行研究の手法では学習に悪影響を及ぼさない場合にも細分化を行い，不要な知識を作成・記憶していたと考える．したがって本研究では，認識した状態が不完全知覚であり，かつ学習に悪影響を及ぼす場合のみ細分化を行う．これにより不要な細分化を防ぎ，知識量の抑制を図る．

さらに，細分化を行う対象について，ロボットが認識状態を経験した回数にも注目する必要があると考える．この理由として，学習においては，ある状態に適した行動を学習するために，その状態で試行錯誤を繰り返す必要があることが挙げられる．また，動的な環境を考えると，あまり起こらない変化により不完全知覚が起きている場合，その状態について細分化を行っても，その知識をあまり利用しないことが挙げられる．したがって，ある程度の回数認識した状態のみを分割することで，学習に合わせて状態知識を増加させることが出来ると考える．

以上のことから，提案手法では経験情報として，認識状態の経験回数，選択する行動，選択した行動によって遷移する認識状態に注目する．そしてこれらの経験情報を基に確率的に細分化を行うことで，知識量の抑制を図り，先行研究における問題を解決する(図 1.6)．

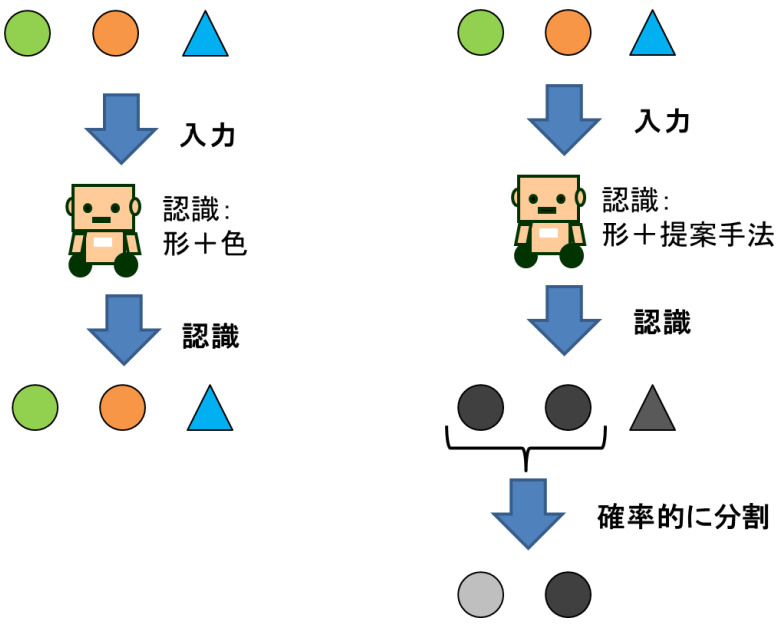

図 1.6　本研究のアプローチ

1.6　本論文の構成

本論文では第 1 章で，まず，ロボット技術の進歩から学習が必要になったことを説明し，学習への影響が大きい，ロボットの状態認識と不完全知覚について説明した．さらに，不完全知覚の改善を図った先行研究を挙げ，先行研究の問題について述べた．そして，先行研究の問題を解決するために，本論文では先行研究の細分化に注目することを述べ，最後に，先行研究の問題点から本研究の目的を述べ，目的達成のアプローチを示した．

第 2 章では不完全知覚について説明する．まずロボットの状態認識について説明し，不完全知覚がどのようなものか説明する．次に学習における環境モデルについて説明し，最後に不完全知覚の学習への影響と，どのような場合に学習に悪影響を及ぼすか説明する．

第 3 章では，先行研究について説明する．まず先行研究がロボット自身の経験情報に注目して不完全知覚の改善を図った研究であることを述べ，どのように経験情報を利用するか説明する．次に，先行研究の手法を対象とする学習ロボットに適用する際のシステムについて説明し，先行研究の概要を説明する．そして，先行研究のシステムと流れ，先行研究で用いる知識について説明し最後に，先行研究の問題を説明する．

第 4 章では，提案手法について説明する．まず提案手法が先行研究の手法をどのように改善したものであるか説明する．そして，提案手法の概要を説明し，提案手法で用いる経験情報について説明する．次に提案手法のシステムと流れについて，先行研究の手法を改善した部分のみ説明する．

第 5 章では，提案手法の不完全知覚に対する有効性を検証するシミュレーション実験について説明する．まず，シミュレーション実験の目的を述べ，シミュレーション実験について説明する．そして実験の結果について，完全知覚，不完全知覚，先行研究，提案手法の比較を行い，考察する．

第 6 章では，本論文全体のまとめを述べ，今後の課題について述べる．

第 2 章　不完全知覚

本章では，まずセンサを用いたロボットの状態認識について説明する．次にセンサを用いた状態認識の問題として知られる不完全知覚とその問題について説明し，そして学習における環境のモデルについて説明する．最後に，不完全知覚が学習に及ぼす影響について説明する．

2.1　センサを用いたロボットの状態認識

ロボットがセンサを用いて環境状態を観測して認識した状態を観測状態と呼び，ロボットが認識し得るn_c個の観測状態の集合を$\mathrm{C} \coloneqq \{c_i | \mathrm{i} = 1,2,\cdots,n_c\}$とする．このとき，ある観測状態$c_i \in \mathrm{C}$を$\mathrm{n}_v$個のセンサからの各入力を$v_j(\mathrm{j} = 1, 2, \cdots, \mathrm{n}_v)$とし，式(2.1)で表す．

$$c_i = \left(v_1, v_2, \cdots, v_{n_v}\right) \tag{2.1}$$

観測状態c_iはセンサの個数n_vが増加したり，v_jの分解能が向上したりすることで，より細かく認識することが出来る．観測状態c_iを細かく認識することで，ロボットはより詳細に状態を認識することが出来るというメリットが存在する．しかし一方で学習において，観測状態c_iのとり得る数が増加することで，学習にかかる時間が増加するというデメリットがある．このため，多くのセンサや，分解能が高いセンサをロボットに搭載することは，必ずしも良いことであるとは限らない．以上のことから，ロボットが作業を行う上で必要な種類，能力のセンサのみを搭載することが望ましいと考えられる．しかし，近年ではロボットが利用される環境は複雑になっており，ロボットに必要なセンサを事前に想定することは困難になっている．

2.2　ロボットの認識能力の不足による不完全知覚の発生

ある時刻 t における環境状態を$s_t \in \mathrm{S}$(Sは環境のとり得る状態の集合)，観測状態を$o_t \in C$と表す．これらの間には一般に，式(2.2)の関係が成り立つ．これは複雑な環境において，環境状態と観測状態は一致しないことを表している．しかし，環境状態と観測状態の間に関係性がないということではない．o_tはセンサを用いてs_tを観測して認識した状態である．したがって，s_tとo_tの間には何らかの関係性が存在している．ただしこの関係性はノイズが含まれた観測状態には存在しない可能性が高い．

$$s_t \neq o_t \tag{2.2}$$

以上のように，センサを用いた状態認識において，ロボットは搭載しているセンサの能力や種類に応じた精度で，環境状態を認識している．このとき，環境状態と観測状態の不一致により，ロボットは異なる複数の環境状態を，同じ観測状態であると認識してしまう可能性がある．このようにある観測状態が異なる複数の環境状態を混同し，それらを区別出来ない状態認識を「不完

全知覚」と呼ぶ．本論文ではこの不完全知覚によって起きる問題に注目する．不完全知覚による問題は強化学習において頻繁に取り上げられている．先行研究及び本論文でも強化学習に基づいて学習と不完全知覚の関係に注目する．また本論文において，環境状態と観測状態が一致する場合，その状態認識を「完全知覚」と呼ぶ．つまり完全知覚においては，環境状態の集合S と観測状態の集合C の間に式(2.3)の関係が成り立ち，ロボットが環境状態を完全に認識出来ることを表している．

$$S = C \tag{2.3}$$

2.3 学習における環境モデル

本論文では実機で用いられることも多い強化学習に基づいて，学習と不完全知覚の関係について論じる．強化学習においては多くの場合，ロボットが行動する環境をマルコフ決定過程(Markov Decision Process , MDP)によりモデル化する． MDP の枠組みではロボットは全ての環境状態を完全に認識できると仮定している．しかし，複雑な環境において一般に，センサを用いた状態認識は不完全性が存在する．このような環境は部分観測マルコフ決定過程(Patially Observable Markov Decision Process, POMDP)によりモデル化する．POMDP は MDP のモデルを拡張し，ロボットの状態認識に不完全性を付加したモデルである．本論文で扱う環境は POMDP に当たり，環境そのものは変化のない静的な環境を扱う．

2.3.1 マルコフ決定過程(MDP)

本論文で注目する強化学習においては多くの場合，ロボットが行動する環境をマルコフ決定過程(MDP)によってモデル化する．マルコフ決定過程はロボットが完全に環境を認識出来る場合に用いることが出来る．マルコフ決定過程ではある時刻において状態s_tで行動$a_t \in M$(Mはロボットがとり得るn_m個の行動の集合で$M \coloneqq \{m_j | j = 1, \cdots, n_m\}$とする)をとったとき，その行動の結果遷移する状態 s_{t+1}と，行動の結果得られる報酬r_tがそれぞれ，s_tとa_tのみに依存した確率に基づいて決定する．また，ロボットは MDP 環境において，強化学習によって最適な政策を得ることが出来る．

2.3.2 部分観測マルコフ決定過程(POMDP)

複雑な環境において，ロボットがセンサを用いて状態認識を行う際には 2.2 節で述べたように，状態認識に不完全性が存在し得る．MDP では環境を完全に認識出来ることを仮定するため，状態認識が不完全知覚である場合においては MDP を用いることは不適切である．このような場合には部分観測マルコフ過程(POMDP)を用いる．POMDP は MDP のモデルを拡張し，ロボットの状態認識に不完全性を付加したモデルである．本論文では状態認識が不完全知覚である場合に注

目するため，POMDP によって環境をモデル化する．

2.4　不完全知覚が学習に及ぼす影響

強化学習において，エージェントはセンサを用いて認識した状態に対して学習を行う．すなわち，エージェントは各観測状態に対し，適した行動を学習していく．観測状態が環境状態と一致する場合，つまり MDP 環境においては，強化学習により適した行動を学習出来ることが知られている．しかし，観測状態が環境状態と一致しない場合，つまり POMDP 環境においては，強化学習により適した行動を学習出来るとは限らない．

不完全知覚が学習に影響を及ぼす可能性について，不完全知覚が起きている観測状態においてとるべき行動に注目する．強化学習において，エージェントは 1 つの観測状態に対して 1 つの適した行動を学習する．したがって，不完全知覚である観測状態において複数の適した行動が存在している場合，その観測状態で適した行動が学習出来なくなると考えられる．学習において，不完全知覚により起こる問題としてループへの落ち込みが知られている．これは，不完全知覚である観測状態で学習した行動と，他の観測状態で学習した行動との間で齟齬が生じ，この 2 状態を往復することがあるという問題である．例として，図 2.1 のような環境において，強化学習を用いて迷路問題を行う場合を考える．エージェントはセンサを用いて，現在のマスの上下左右の壁の有無によって状態を認識するとする．このような状態認識において，同じ色で示したマスでは同じ観測状態が得られるため，これらの観測状態は不完全知覚が起きている観測状態である．エージェントは図 2.1 の右図に示した行動を学習することで，スタートからゴールまで無駄なく移動が行えるようになる．このとき，■で示したマスにおいて適した行動が，環境の上部では左への移動，下部では右への移動であり，エージェントはどちらかの行動しか学習出来ない．ここで右への移動を学習したとすると，上部の■でも右へ移動する．この場合エージェントは■で左へ移動，■で右へ移動するため，■と■のマスで往復が起きる．このように不完全知覚によって，適した行動が学習出来た観測状態と学習出来なかった観測状態の間でループに陥る場合があり，作業が行えなくなってしまうという問題がある．

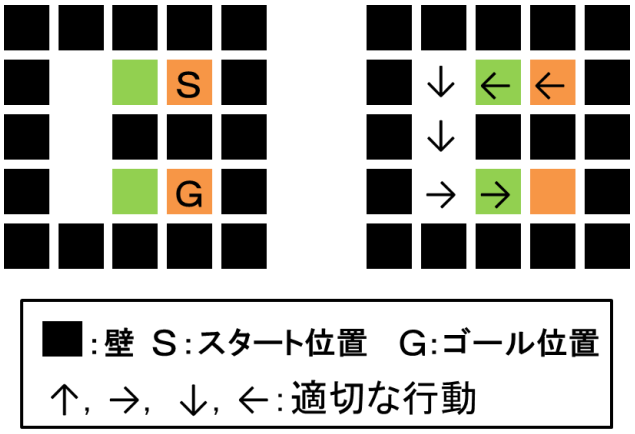

図 2.1　不完全知覚が起きる迷路環境

一方，複数の環境状態において，各環境状態で適した行動が同じ場合には，学習に悪影響を及ぼす可能性は低いと考えられる．本研究では，このような場合には不完全知覚によって混同している環境状態を区別する必要は無いと考える．また，このとき複数の環境状態において適した行動を一度に獲得できる可能性がある．

第 3 章　先行研究

2 章では不完全知覚について説明し，不完全知覚が学習にどのような影響を与えるか説明した．3 章ではまず，先行研究[10]が不完全知覚を改善するために注目した点について述べ，その上で，先行研究がどのように不完全知覚を改善するかを説明する．次に，先行研究が提案した手法を説明し，最後に，先行研究の手法によって不完全知覚を改善する際に存在する問題を挙げる．ここで，本論文では先行研究に経験知識を修正する方法を追加したものを比較対象とする．本章で説明するのも，この経験知識の修正を追加したものである．

本章でも 2 章で定義したように，ロボットが認識し得るn_c個の観測状態の集合を$\mathrm{C} \coloneqq \{c_i|\mathrm{i} = 1,2,\cdots,n_c\}$，ロボットがとり得る$n_m$個の行動の集合を$\mathrm{M} \coloneqq \{m_j|j = 0,1,\cdots,n_m\}$とし，ある時刻tにおける観測状態を$o_t \in C$，それに対しとった行動を$a_t \in \mathrm{M}$とする．

3.1　経験情報を用いた不完全知覚の改善

先行研究では，センサ情報のみを用いて状態認識を行っているために不完全知覚が起こると考えた．そこでセンサ以外の情報として，ロボット自身の経験情報に注目している．先行研究の提案手法では，ロボットはセンサを用いて認識した観測状態が不完全知覚であるかを，自身の経験情報を用いることで判断する．そして，観測状態が不完全知覚により複数の環境状態を混同していると判断したとき，その観測状態を，不完全知覚である観測状態そのものと，直前の観測状態と行動を合わせた新たな状態の 2 つに細分化する(図 3.1)．そして，新たな状態を「状態知識」として記憶し，その後の状態認識に用いる．ロボットは状態知識も状態認識に利用することで，不完全知覚である観測状態を，不完全知覚である観測状態そのものと，ある状態遷移の結果得られた観測状態としてそれぞれ区別することが出来る．

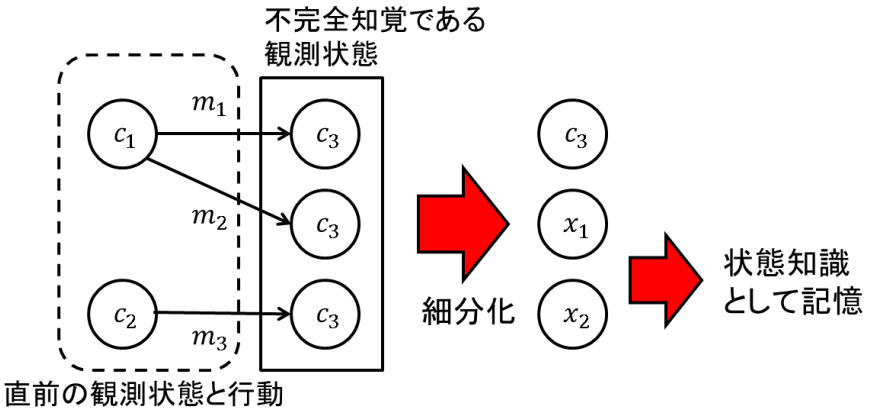

図 3.1　経験情報を用いた観測状態の細分化

3.2　先行研究の概要

まず，先行研究の提案手法を適用したロボットの学習について説明する．先行研究はセンサを用いて状態を認識し，強化学習を行うロボットを対象としている．対象ロボットに先行研究の提案手法を適用した場合，最終的には先行研究の提案手法により決定した状態に基づいて強化学習を行う．この状態を認識状態と呼び，ロボットが認識し得る$n_{\hat{c}}$個の認識状態の集合を$\hat{C} \coloneqq \{\hat{c}_k | k = 0,1,\cdots,n_{\hat{c}}\}$とする．また，時刻tにおける認識状態を$\hat{o}_t \in \hat{C}$とする．このとき，先行研究における提案手法と強化学習の関係を図 3.2 に示す．先行研究の提案手法を適用したロボットはセンサを用いて環境状態を観測し，センサ値から観測状態を認識する．さらに，提案手法により認識状態を決定し，認識状態に対して学習を行う．このシステムにおいて，先行研究の提案手法ではロボットは行動を重ねることで状態知識を獲得し，認識出来る状態が増加する．そのため，強化学習の学習空間も認識出来る状態の増加に合わせて拡大する必要がある(図 3.3)．このとき，増加した認識状態に対応した学習空間(図 3.3 の黄緑色の部分)に与える初期値は，最初に学習空間を構成した際と同様の初期値を与える．したがって，新たに増加した認識状態でも試行錯誤を繰り返すことで学習を行っていく必要がある．

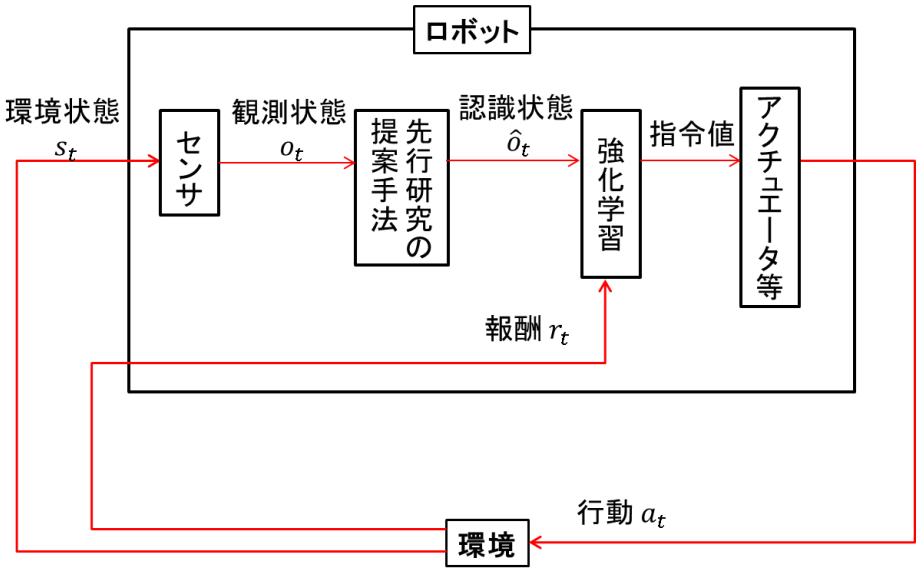

図 3.2　先行研究の提案手法と強化学習の関係

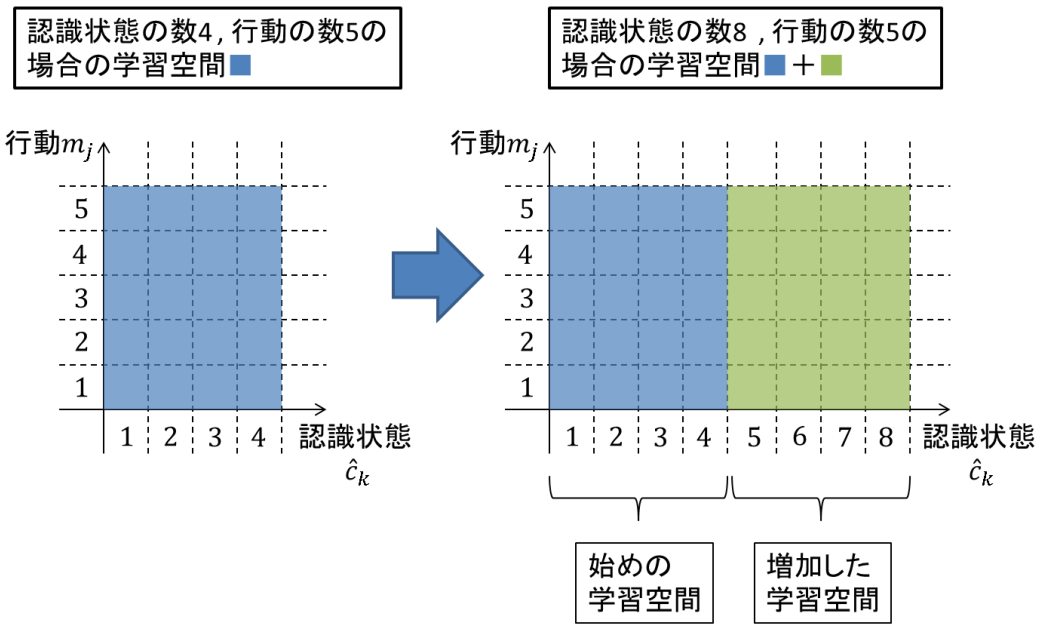

図 3.3　認識し得る状態の増加に伴う学習空間の拡大

次に，先行研究の手法で行う状態認識と構成要素について説明する．先行研究の手法では，ロボットは自身の経験情報を用いて認識状態が不完全知覚か判断し，不完全知覚である場合にその認識状態を細分化する．細分化によって状態知識を記憶し，それを状態認識に用いることでより詳細に状態を認識し，不完全知覚の改善を図っている．先行研究の提案手法は大きく分けて以下の 4 つのモジュールから構成される．

・状態認識部

・不完全知覚判定部

・細分化部

・経験情報蓄積部

さらに，先行研究の手法ではロボットは以下の 2 つの知識を持つ．

・経験知識E

・状態知識X

経験知識Eはエージェント自身の過去の経験の集合であり，状態知識Xは細分化によって得た，不完全知覚である任意の認識状態$\hat{c}_k \in \hat{C}$に関する知識の集合である．経験知識は「経験情報蓄積部」において作成され，「不完全知覚判定部」である時刻tにおいて，その直前の認識状態$\hat{o}_{t-1} \in \hat{C}$が不完全知覚か判断するために用いられる．また，状態知識は不完全知覚と判断された認識状態$\hat{o}_{t-1}$を「細分化部」で細分化することで作成され，「状態認識部」において観測状態と合わせて状態認識に用いる．これにより，今まで不完全知覚によって$\hat{o}_{t-1}$と認識していた状態を，新しく別な認識状態$\hat{o}'_{t-1} \in \hat{C}$として認識することが出来るようになる．このとき，それまでに記憶してきた経験知識には，$\hat{o}_{t-1}$で経験した情報と，不完全知覚により$\hat{o}_{t-1}$として認識していた認識状態

$\hat{o}'_{t-1}$で経験した情報が，どちらも$\hat{o}_{t-1}$で経験した情報として記憶されていることになる．よって，細分化を行う際に$\hat{o}_{t-1}$を含む経験知識を全て忘却することで，経験情報の修正を行っている．これが本章の始めに述べた経験情報の修正である．これらのモジュールと知識を用いた先行研究の提案手法の概略図を図 3.4 に示す．

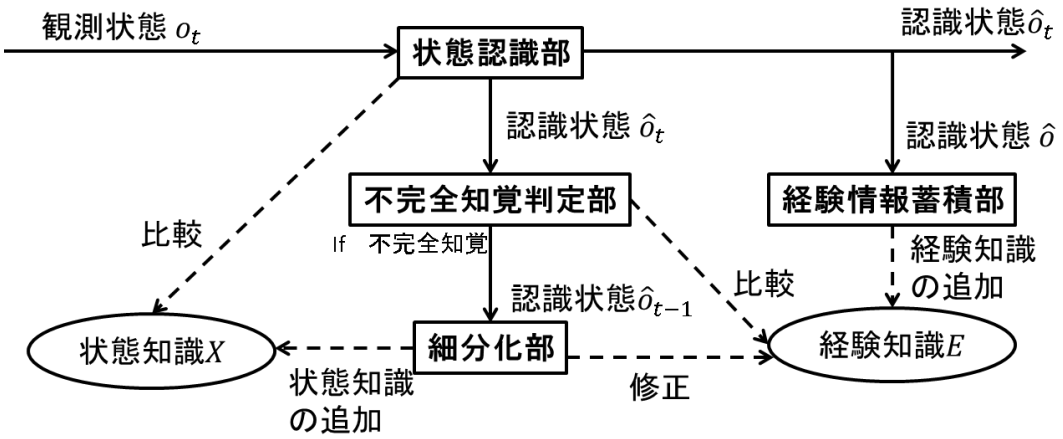

図 3.4　先行研究の提案手法の概略図

3.3　先行研究の提案手法の流れ

これらのモジュールと知識を用いた先行研究の提案手法において，細分化の対象となるのは直前の時点t − 1における認識状態$\hat{o}_{t-1} \in \hat{C}$である．これは，不完全知覚判定部において不完全知覚か判定を行う際，対象となる認識状態とそれに対する行動，その結果遷移した次時点の認識状態を用いるためである．現時点ではまだロボットが行動を決定していないものとするため，不完全知覚か判定する情報が足りず，不完全知覚か判定出来ない(図 3.5)．

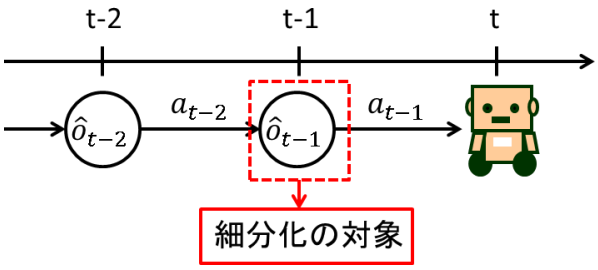

図 3.5　細分化を行う対象

実際に，ある時点 t における先行研究の手法の流れを説明する．ロボットは始めに，状態認識部で現在の観測状態$o_t \in C$と状態知識集合Xから現在の認識状態$\hat{o}_t \in \hat{C}$を決定する．ここで，状態知識集合Xはロボットが記憶しているn_X個の状態知識の集合であり，$X := \{x_i | i = 1,2,\cdots,n_X\}$とする．次に，不完全知覚判定部で経験知識集合Eを用い，直前の認識状態 $\hat{o}_{t-1} \in \hat{C}$ が不完全知覚か判定す

る．ここで，経験知識集合Eはロボットが記憶しているn_E個の経験知識の集合であり，$E \coloneqq \{e_j | j = 1,2,\cdots,n_E\}$とする．ロボットが認識状態$\hat{o}_{t-1}$を不完全知覚であると判断した場合，直前の認識状態$\hat{o}_{t-1}$をさらに1つ前，つまり現時点tから見て2つ前の時点$t-2$における認識状態$\hat{o}_{t-2} \in \hat{C}$と行動$a_{t-2} \in M$を用いて細分化部で細分化する．このとき，経験知識集合Eの認識状態$\hat{o}_{t-1}$を含む全ての経験知識$e_w \in E$を忘却することで，経験知識Eを修正する．不完全知覚でないと判断した場合は認識状態$\hat{o}_{t-1}$をそのまま直前の認識状態$\hat{o}_{t-1}$として決定する．最後に，経験情報蓄積部において，直前の認識状態$\hat{o}_{t-1}$と行動a_{t-1}，現在の認識状態$\hat{o}_t$を用いて経験知識を作成する．先行研究の状態認識から細分化までの流れを図 3.6 にまとめる．

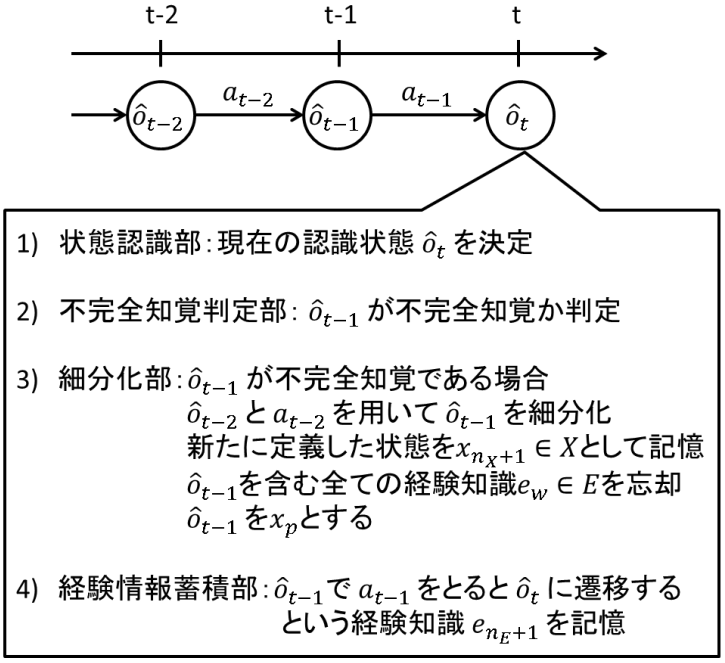

図 3.6　先行研究の手法の流れ

3.4　先行研究で用いていた知識

先行研究では，経験知識と状態知識という 2 つの知識を用いて不完全知覚の改善を図っている．本節ではこの 2 つの知識の定義を行い，それぞれの利用法を説明する．

3.4.1　経験知識

経験知識とはロボット自身の経験を知識化したものである．先行研究ではロボットが得た認識状態と，それに対して選択した行動に注目してロボット自身の経験を知識化する．したがって，ロボットがある時点tで知識化する経験知識はそれまでの経験知識の数をn_Eとすると，$e_{n_E+1} \in E$（Eは経験知識の集合を表す）と表せ，以下の式(3.1)で定義する．式(3.1)において，$\hat{o}_{t-1}, a_{t-1}$は時点tの直前の認識状態とそれに対して選択した行動，$\hat{o}_t$はその結果時点tにおいて得た認識状態を表

している．つまり経験知識は，ある時点で経験した状態遷移を表した知識である．

$$\mathrm{e}_{n_E+1} = (\hat{\mathrm{o}}_{\mathrm{t}-1}, \mathrm{a}_{\mathrm{t}-1}, \hat{\mathrm{o}}_{\mathrm{t}}) \tag{3.1}$$

先行研究ではこの経験知識を用いて，直前の認識状態が不完全知覚か判断する．不完全知覚である認識状態が複数の環境状態を含んでいることから，同じ行動を選択しても異なる認識状態に遷移し得ると考えられる．したがって，ある時点tにおける認識状態に遷移する状態遷移$(\hat{\mathrm{o}}_{\mathrm{t}-1}, \mathrm{a}_{\mathrm{t}-1}, \hat{\mathrm{o}}_{\mathrm{t}})$を経験知識と比較し，同じ認識状態と行動に対し異なる認識状態に遷移している経験知識が存在する場合に，直前の認識状態を不完全知覚であると判断する．

3.4.2 状態知識

状態知識とは不完全知覚である認識状態を細分化することで得られる知識である．具体的には，不完全知覚である認識状態が，どのような状態遷移によって得られたかを表している．この状態知識を状態認識に用いることで，より細かな状態認識を行うことが可能になる．ロボットがある時点tで知識化する状態知識はそれまでの状態知識の数をn_Xとすると$\mathrm{x}_{n_X+1} \in \mathrm{X}$と表せ，以下の式(3.2)で定義する．式(3.2)において，$\hat{\mathrm{o}}_{t-1}$は細分化の対象である認識状態，$\hat{\mathrm{o}}_{t-2}$はその直前の認識状態，a_{t-2}は$\hat{\mathrm{o}}_{t-2}$に対して選択した行動を表している．

$$x_{n_X+1} = (\hat{o}_{t-1}, \hat{o}_{t-2}, a_{t-2}) \tag{3.2}$$

先行研究ではこの状態知識を用いて，状態認識部でより詳細に状態を認識する．現在の観測状態に直前の認識状態と行動を合わせて状態知識と比較し，一致する知識があった場合は現在の観測状態はその状態遷移によって得られる観測状態として個別に認識することができる．

3.5 状態認識部

本節では状態認識部で認識状態を決定する具体的なアルゴリズムについて説明する．状態認識部ではある時点tにおいて，現在の認識状態$\hat{o}_t$を現在の観測状態$o_t \in C$と状態知識Xを用いて決定する(図 3.7)．このとき，直前の認識状態$\hat{o}_{t-1} \in \hat{C}$と行動$a_{t-1} \in M$も用いる．具体的には観測状態 o_tに，直前の認識状態 $\hat{o}_{t-1}$ で行動 a_{t-1}を選択した結果遷移した場合を表すある状態知識 $x_p \in \mathrm{X}$を検索する．一致する状態知識が見つかればx_pを現在の認識状態として決定する．一致する状態知識が見つからなかった場合は，現在の観測状態をそのまま認識状態として決定する．以下に状態認識部の処理の流れを示す．

1) 現在の観測状態をo_t，直前の認識状態と行動をそれぞれ$\hat{o}_{t-1}, a_{t-1}$とする．

2) 状態知識集合 X の中に$(o_t, \hat{o}_{t-1}, a_{t-1})$となる状態知識$x_p \in \mathrm{X}$が存在するか検索．

3) $\exists x_p \in \mathrm{X}$($x_p$が存在する)場合，現在の認識状態 $\hat{o}_t$ をx_pとして決定する．

4) $\nexists x_p \in X$(x_pが存在しない)場合，現在の認識状態 $\hat{o}_t$を o_t として決定する．

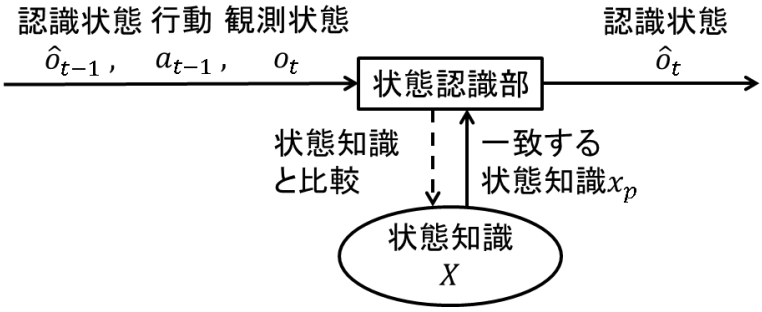

図 3.7　状態認識部の概要

3.6　不完全知覚判定部

本節では不完全知覚判定部で不完全知覚かどうか判断する具体的なアルゴリズムについて説明する．不完全知覚判定部ではある時点tにおいて，直前の認識状態$\hat{o}_{t-1} \in \hat{C}$が不完全知覚かどうか経験知識E を用いて判断する(図 3.8)．このとき，現在の認識状態$\hat{o}_t \in \hat{C}$，直前に選択した行動$a_{t-1} \in \mathrm{M}$も用いる．具体的には，認識状態 $\hat{o}_{t-1}$ で行動 a_{t-1} を選択したときの経験知識を検索し，該当したある経験知識$e_q \in \mathrm{E}$において，その結果遷移した認識状態を$\hat{o}_t$と比較する．もし，遷移した認識状態が異なる経験知識$e_q \in \mathrm{E}$が存在する場合は，認識状態$\hat{o}_{t-1}$を不完全知覚と判断し，細分化部で細分化を行う．その他の場合においては不完全知覚ではないと判断し，細分化は行わない．つまり，細分化部を利用するのはこの不完全知覚判定部で，認識状態$\hat{o}_{t-1}$が不完全知覚であると判断された場合のみである．以下に不完全知覚判定部の処理の流れを示す．以下の処理において$\hat{c}_k$は任意の認識状態($\hat{c}_k \in \hat{C}$)を表す．

1) 時点tにおける認識状態を$\hat{o}_t$とし，その直前の認識状態と行動をそれぞれ$\hat{o}_{t-1}, a_{t-1}$とする．

2) 経験知識集合Eの中から，直前の認識状態と行動に関する知識$(\hat{o}_{t-1}, a_{t-1}, \hat{c}_k)$ が存在するか検索．

3) 見つかった経験知識を$(\hat{o}_{t-1}, a_{t-1}, \hat{c}'_k)$とする．

4) $\hat{c}'_k \neq \hat{o}_t$ の場合，$\hat{o}_t$ を不完全知覚と判断して細分化を行う．

5) $\hat{c}'_k = \hat{o}_t$ の場合，経験知識集合E内の全ての知識と比較するまで 2)に戻る．

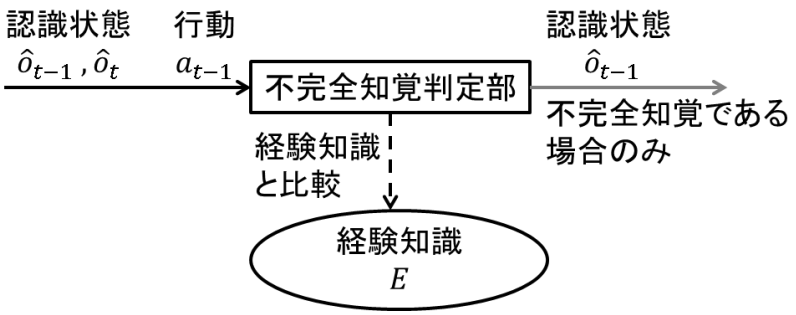

図 3.8　不完全知覚判定部の概要

3.7　細分化部

細分化部ではある時点 t において，不完全知覚と判断された認識状態$\hat{o}_{t-1} \in \hat{C}$を細分化する(図 3.9)．そのために，認識状態$\hat{o}_{t-1}$とその直前，つまり現在からみて2つ前の時点$t-2$における認識状態$\hat{o}_{t-2} \in \hat{C}$と行動$a_{t-2} \in M$を用いる．具体的に，時点tまでに認識し得た認識状態の数をn_Xとし，細分化によって認識状態 $\hat{o}_{t-1}$ を元の認識状態$\hat{o}_{t-1}$と，新たな状態x_{n_X+1}に分割する．このとき，状態知識$x_{n_X+1} = (\hat{o}_{t-1}, \hat{o}_{t-2}, a_{t-2})$と定義し，ロボットは$x_{n_X+1} \in \mathrm{X}$として状態知識$x_{n_X+1}$を記憶することで，状態認識に利用することが可能になる．このとき，経験知識集合E には新たな状態x_{n_X+1} を認識状態$\hat{o}_{t-1}$として認識していたときの経験知識が残っている．しかし，ロボットはそれらを新たな状態x_{n_X+1}で経験した情報として修正することは出来ない．そこで，$\hat{o}_{t-1}$を含む経験知識$e_w \in \mathrm{E}$を全て忘却することで，経験知識集合E に齟齬が生じないよう修正する．以下に細分化部の処理の流れを示す．以下の処理でm_jは任意の行動($m_j \in \mathrm{M}$)を，$\hat{c}_k$は認識状態$\hat{o}_{t-1}$で各行動m_j をとった時に遷移し得る各認識状態($\hat{c}_k \in \hat{C}$)をそれぞれ表している．

1) 細分化を行う認識状態を$\hat{o}_{t-1}$とし，その直前の認識状態と行動をそれぞれ$\hat{o}_{t-2}$, a_{t-2} とする
2) $\hat{o}_{t-1}$を$\hat{o}_{t-1}$とx_{n_X+1}に細分化し，新たな状態知識$x_{n_X+1} = (\hat{o}_{t-1}, \hat{o}_{t-2}, a_{t-2})$を定義する
3) $x_{n_X+1} \in \mathrm{X}$として新たな状態知識を記憶する
4) $\hat{o}_{t-1}$を含む経験知識$\mathrm{e}_w = (\hat{o}_{t-1}, m_j, \hat{c}_k)$ と$\mathrm{e}_w = (\hat{c}_k, m_j, \hat{o}_{t-1})$ を忘却することで経験知識集合E を修正する($\forall c_k \in \hat{C}, \forall m_j \in \mathrm{M}$)
5) $\hat{o}_{t-1}$をx_{n_X+1}として決定する

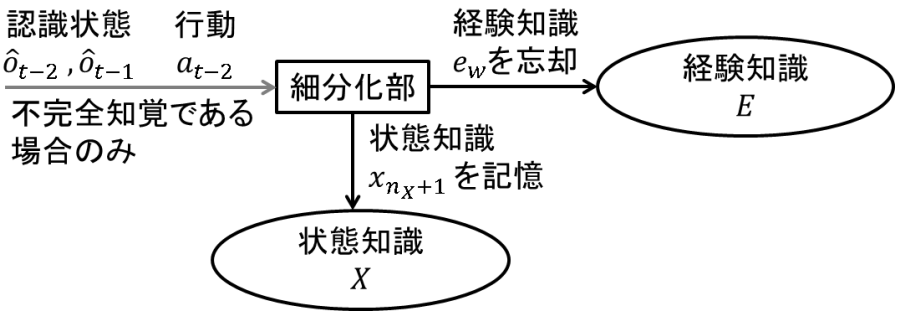

図 3.9　細分化部の概要

3.8　経験情報蓄積部

経験情報蓄積部では，ロボットが経験した状態遷移を経験知識として記憶する(図 3.10)．ある時刻tで記憶する経験知識は，現在の認識状態$\hat{o}_t \in \hat{C}$，直前の認識状態と行動 $\hat{o}_{t-1} \in \hat{C}$, $a_{t-1} \in M$ を用いて知識化する．このとき，時点tまでに記憶していた経験知識の数をn_Eとすると，時点tで新たに知識化した経験知識を$e_{n_E+1} = (\hat{o}_{t-1}, a_{t-1}, \hat{o}_t)$と表せ，ロボットは$e_{n_E} \in$ E として記憶することで不完全知覚の判断に利用することが可能になる．以下に経験情報蓄積部の処理の流れを示す．

1) 現在の認識状態を$\hat{o}_t$とし，その直前の認識状態と行動をそれぞれ$\hat{o}_{t-1}, a_{t-1}$とする．
2) 新たに知識化する経験知識を$e_{n_E+1} = (\hat{o}_{t-1}, a_{t-1}, \hat{o}_t)$として定義する．
3) $e_{n_E+1} \in$ E として，新たな経験知識を記憶する．

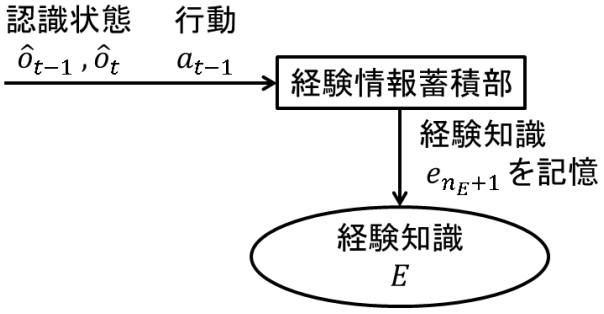

図 3.10　経験情報蓄積部の処理の流れ

3.9　先行研究の提案手法の問題点

先行研究ではシミュレーション実験により提案手法の有効性を検証していた．実験結果から，状態知識と経験知識を利用することで不完全知覚を改善し，学習が可能になることが示されている．しかしその一方で，知識量の増加という問題が発生している．知識量の増加が問題になる理由の 1 つとして，ロボットのメモリ量における問題がある．これは，先行研究においてロボットが行動を重ねることで無限に知識が増加するが，ロボットのメモリ量は有限であるため，増加す

る知識を全て記憶することは出来ないという問題である．また，他の理由として学習における問題がある．これは，状態知識が増加することで認識出来る状態が増加し，学習にかかる時間が増加するという問題である．ロボットは増加した認識状態に対し，新たに試行錯誤を繰り返して適した行動を学習する必要がある．そのため，増加した認識状態に対して適した行動を学習するまでに時間がかかると考えられる．また，急激に認識出来る状態が増加した場合には学習が追い付かず，環境内の学習が済んでいない認識状態の割合が増加すると考えられる．このとき学習において，ロボットは学習が済んでいない認識状態に遷移したとき，行動を評価することが困難になる．それによって学習が遅れることが考えられる．以上のことから，本研究では知識量の増加を先行研究の問題点であると考える．そこで今回，本研究では先行研究の細分化に注目する．そして先行研究の手法を改善することで知識量を抑制しつつ，強化学習における不完全知覚問題を解決する手法を提案する．

第 4 章　提案手法

先行研究ではロボット自身の経験情報に注目し，状態知識と経験知識を作成・利用することで不完全知覚を改善する手法を提案していた. しかし, 不完全知覚を改善することが出来た一方で, 知識量の増加という問題が新たに発生していた. 本章ではまず, 知識量の増加という問題に対し, 本研究が細部化をどのように行うことで知識量を抑制するか述べる．次に，先行研究の手法をどのように改善するか説明する．最後に，先行研究の手法を改善した，本研究の手法を提案する．

本章でも 2 章や 3 章で定義したように，ロボットが認識し得るn_c個の観測状態の集合を$\mathrm{C} \coloneqq \{c_i|\mathrm{i} = 1,2,\cdots,n_c\}$，ロボットがとり得る$n_m$個の行動の集合を$\mathrm{M} \coloneqq \{m_j|j = 0,1,\cdots,n_m\}$とし，ある時点tにおける観測状態を$o_t \in C$，それに対しとった行動を$a_t \in M$とする．また，3 章で定義したように，ロボットが認識し得る$n_{\hat{c}}$個の認識状態の集合を$\hat{C} \coloneqq \{\hat{c}_k|k = 1,2,\cdots,n_{\hat{c}}\}$とし，ある時点tにおける認識状態を$\hat{o}_t \in \hat{C}$とする．

4.1　知識量の抑制による先行研究の問題解決

先行研究はセンサ情報に加え，ロボット自身の経験情報を状態認識に用いることで不完全知覚を改善した．その一方で，知識量の増加という問題が存在する．そこで，本研究では状態の細分化に注目する．具体的には，細分化を行う対象と方法に注目する．

先行研究では，不完全知覚により複数の環境状態を混同している認識状態を，確定的に細分化していた．このとき，複数の環境状態を混同している認識状態において，2.5 節で述べたように，混同している環境状態で適した行動が一致していれば，その認識状態で適した行動を学習出来る可能性は高い．したがって，その認識状態で混同している環境状態で細分化を行う必要は無いと考える．また，細分化を行う必要があると判断した認識状態で，確定的に細分化を行うことで，頻繁に細分化を行うことがあると考えられる．これにより知識量が急激に増加すると考えられる．以上のことから，本研究では先行研究のような確定的な細分化が先行研究の問題の原因であると考える．そこで，本研究では細分化を行う確率をロボットの経験情報を基に決定し，確率的に細分化を行うことを考える．細分化を確率的に行うことで細分化を抑制し，知識量の増加を抑制する．更に，その認識状態で行動が学習出来ているかどうか，複数の環境状態が混同されているかどうかに注目することで，学習に悪影響を及ぼしている認識状態のみを判断し，不要な知識の増加を抑制することが出来ると考える．また，行動が学習出来ているかを判断するために，認識状態を経験した回数にも注目し，学習初期のような状況を不完全知覚による悪影響がある状況と区別する．

以上より，本研究では，複数の環境状態を混同しているために適した行動が複数あり，ある程度経験しても適した行動を学習出来なかった認識状態に対し，確率的に細分化を行う．このよう

な認識状態を判断するため，本研究ではロボットの過去の経験情報として，「判断する認識状態の認識回数」（以下，認識回数），「判断する認識状態で選択した行動の分散」（以下，選択した行動の分散），「判断する認識状態で選択した行動の結果，遷移した認識状態の分散」（以下，遷移した認識状態の分散）の 3 つに注目する．これらはそれぞれ，認識状態をどれくらい経験したか，適した行動が複数あるか，複数の環境状態を混同しているかを判断出来ると考えられるものである．提案する手法では，これらの経験情報を用いて細分化を行う確率を決定し，その確率に応じて細分化を行う．提案手法で行う細分化の判定アプローチを図 4.1 に示す．

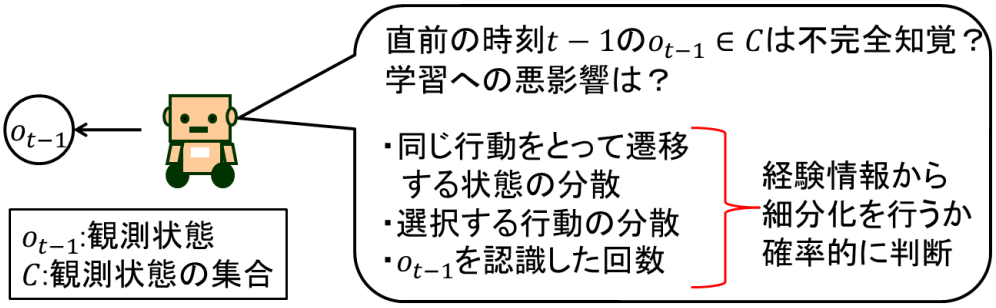

図 4.1　提案手法における細分化の判定アプローチ

また，これらを算出し得る状態遷移の経験回数を経験知識として記憶し，細分化の判断を行うときのみ 3 つの経験情報を算出，利用する．これにより，本研究の提案手法を適用したロボットと，先行研究の提案手法を適用したロボットは，どちらも経験した状態遷移の情報を経験知識として記憶するため，同じ状態遷移を経験すれば記憶する経験知識の数は同数になると考えられる．したがって本研究において，先行研究と知識量を比較する際には，状態知識のみに注目して比較する．

4.2　改善した手法による状態認識

本研究では，先行研究の手法の一部を変更することで改善し，4.1 節で述べた確率的な細分化を行う手法を提案する．具体的な変更点としてはまず，細分化を行うかどうかの判断を変更した．先行研究では，細分化の対象となる認識状態と，過去にロボット自身が経験した状態遷移から不完全知覚であるかを判断して確定的に細分化を行っていた．この細分化の判断を，本研究では細分化の対象となる認識状態の経験回数，選択した行動の分散，遷移した認識状態の分散の 3 つの経験情報を用いて細分化を行う確率を決定し，決定した確率に基づいて確率的に細分化を行うように変更した．また，細分化の判断に用いる情報の変更に伴い，先行研究では経験した状態遷移から経験知識を知識化していた「経験情報蓄積部」を，経験した状態遷移に加えてその状態遷移

の経験回数から経験知識を知識化するように変更した．他のモジュールには変更を加えていない．そのため，先行研究の手法と同様に，強化学習を行うロボットに適用可能である．対象のロボットに本研究の提案手法を適用した場合のシステム概略図を図 4.2 に示す．

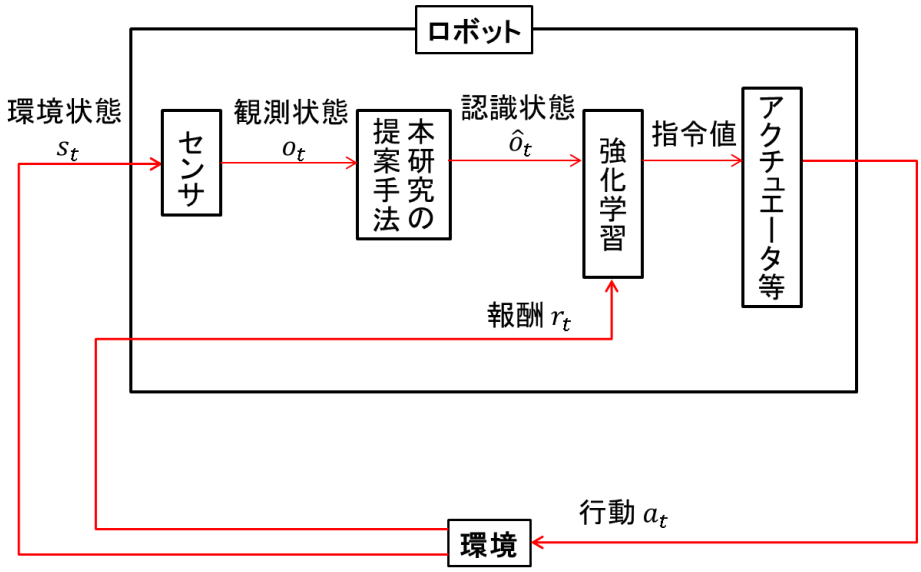

図 4.2　本研究の提案手法を適用したシステムの概略図

本研究の提案手法では，ロボットは自身の経験情報を用いて，認識状態を細分化する確率を決定し，それに基づいて細分化を行う．細分化によって先行研究と同様の状態知識を記憶し，状態認識に用いることで不完全知覚の解決を図る．本研究で提案する手法は以下の 4 つのモジュールから構成される．

・状態認識部

・細分化判定部

・細分化部

・経験情報蓄積部

「状態認識部」,「細分化部」は先行研究の手法と同様のものである．「細分化判定部」は先行研究の手法における「不完全知覚判定部」を改善したものであり，「経験情報蓄積部」で知識化した経験知識を用いて 4.1 節で述べた 3 つの経験情報を算出し，認識状態を細分化する確率を決定，決定した確率に基づき細分化を行うか判断する．「経験情報蓄積部」では状態遷移とその遷移を経験した回数を知識化し，経験知識として記憶する．先行研究の手法と同様，提案手法においてロボットは以下の 2 つの知識を持つ．

・経験知識 E

・状態知識 X

状態知識は先行研究と同様の情報を扱う．経験知識は先行研究の遷移前後の認識状態と遷移前の

行動に，その状態遷移の経験回数を合わせて知識化し，経験知識として記憶している．これら 4 つのモジュールと 2 つの知識を用いた本研究の提案手法の概略図を図 4.3 に示す.

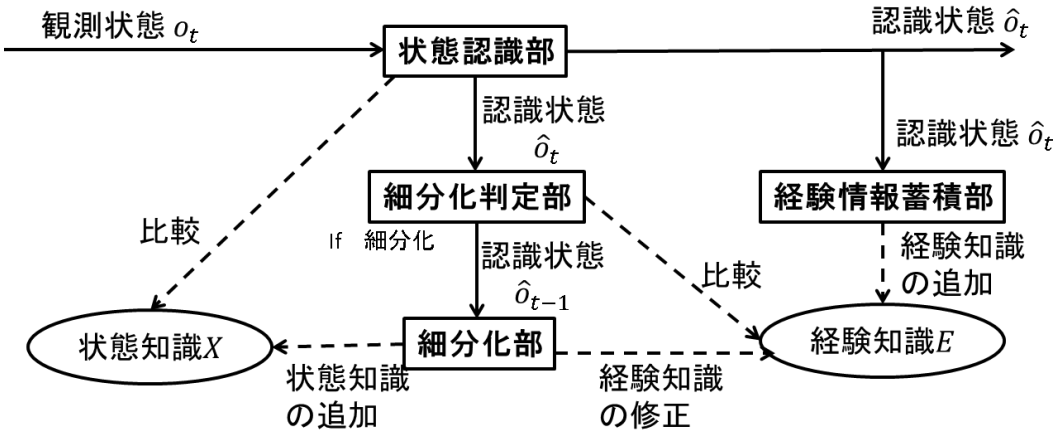

図 4.3　本研究の提案手法の概略図

4.3　提案手法の流れ

本研究で提案する手法は 4.2 節で挙げた 4 つのモジュールと 2 つの知識を用いる．提案手法において，細分化を行うか判断する対象は先行研究と同様の理由から，ロボットから見て直前の時点における認識状態 $\hat{o}_{t-1} \in \hat{C}$ である．実際に，ある時点 t における本研究の提案手法の流れを説明する.

ロボットは始めに，状態認識部で現在の観測状態$o_t \in C$と状態知識 X から現在の認識状態 $\hat{o}_t \in \hat{C}$ を決定する．ここで，状態知識Xはロボットが記憶しているn_X個の状態知識の集合であり，$X \coloneqq \{x_i | i = 1,2,\cdots,n_X\}$とする．次に，細分化判定部で経験知識集合Eを用いて直前の認識状態 $\hat{o}_{t-1}$ を細分化する確率を決定し，決定した確率に基づき細分化を行うか判断する．経験知識集合E はロボットが記憶しているn_E個の経験知識の集合であり，$E \coloneqq \{e_j | j = 1,2,\cdots,n_E\}$とする．ロボットが認識状態 $\hat{o}_{t-1}$ で細分化を行うと判断した場合，直前の認識状態 $\hat{o}_{t-1}$ をさらに 1 時点前，つまりロボットから見て 2 時点前の認識状態 $\hat{o}_{t-2} \in \hat{C}$と行動$a_{t-2} \in M$を用いて細分化部で細分化する．このとき先行研究と同様に，$\hat{o}_{t-1}$を細分化した場合には$\hat{o}_{t-1}$を含む経験知識e_wを忘却することで経験知識集合Eの修正を行う．細分化を行わない場合には，認識$\hat{o}_{t-1}$をそのまま直前の認識状態として決定する．最後に，経験情報蓄積部において，もし直前の認識$\hat{o}_{t-1}$と行動 a_{t-1}，現在の認識$\hat{o}_t$，その状態遷移の経験回数$N(\hat{o}_{t-1}, a_{t-1}, \hat{o}_t)$からなる経験知識があれば，その状態遷移の経験回数を1回増加し，もし一致する経験知識がなければ過去にこの状態遷移の経験を知識化していないため，状態遷移を経験した回数を1回とし，遷移前の認識状態と行動 $\hat{o}_{t-1}, a_{t-1}$ と遷移後の認識状態$\hat{o}_t$，この状態遷移を経験した回数$N(\hat{o}_{t-1}, a_{t-1}, \hat{o}_t)$を用いて，経験知識を作成して記憶する．以上の流れを図 4.4 にまとめる.

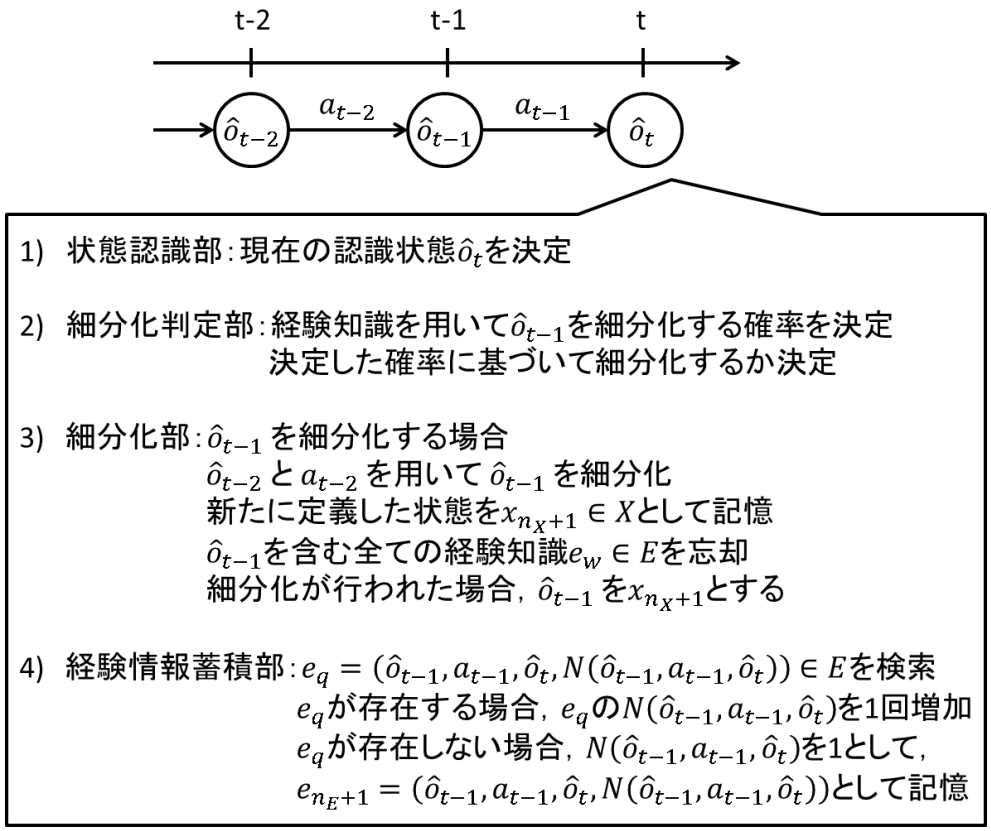

図 4.4　本研究の提案手法の流れ

4.4　経験情報に基づいた確率的な細分化

先行研究において，細分化は「不完全知覚判定部」で認識状態が不完全知覚により複数の認識状態を混同していると判断した場合に，確定的に行っていた．その時に利用していた情報は，ロボットが過去に経験した状態遷移の情報（遷移前の認識状態と行動，遷移後の認識状態）であった．本論文ではこの判断，および確定的な細分化により知識量の増加の問題が発生したと考え，以下の 3 つの経験情報を用いて細分化を行う確率を決定し，決定した確率に基づいて細分化を行うか判断する．

・認識回数
・選択した行動の分散
・遷移した認識状態の分散

これらの経験情報は，経験知識に含まれる状態遷移の経験回数を用いて算出することが出来る．経験知識とこれらの経験情報について，経験知識は 4.4.1 項，認識回数は 4.4.2 項，選択する行動の分散は 4.4.3 項，遷移する認識状態の分散は 4.4.4 項でそれぞれ詳しく説明する．

実際にある時点 t において直前の時点$\mathrm{t}-1$の認識状態$\hat{o}_{t-1} \in \hat{C}$を細分化する確率を決定するまでの流れを説明する．まず，認識状態$\hat{o}_{t-1}$をこれまでに認識した回数$\mathrm{N}(\hat{o}_{t-1})$を式(4.1)で算出する．

$$\mathrm{N}(\hat{o}_{t-1}) = \sum_{j \in u} \sum_{k \in v} N\left(\hat{o}_{t-1}, m_j, \hat{c}_k\right) \tag{4.1}$$

式(4.1)において，m_jはロボットが認識状態$\hat{o}_{t-1}$でとり得る各行動($m_j \in \mathrm{M}$)を，c_kは$\hat{o}_{t-1}$で各行動m_jをとった結果遷移し得る各認識状態($\hat{c}_k \in \hat{C}$)をそれぞれ表す．また，uはロボットが認識状態$\hat{o}_{t-1}$でとり得る全ての行動の行動番号を，vは認識状態$\hat{o}_{t-1}$で各行動m_jをとった際に遷移し得る全ての認識状態の状態番号を表している．つまり$\mathrm{N}(\hat{o}_{t-1})$は$\hat{o}_{t-1}$で経験した全ての状態遷移の経験回数の和である．この$\mathrm{N}(\hat{o}_{t-1}, m_j, c_k)$を用いて経験情報をそれぞれ算出する．次に，認識状態$\hat{o}_{t-1}$ に対して選択した行動の分散を$\sigma_a{}^2(\hat{o}_{t-1})$と表し，4.4.3 項で説明する方法で算出する．そして，認識状態$\hat{o}_{t-1}$と実際に選択した行動 a_{t-1}の結果遷移した認識状態の分散を$\sigma_{\hat{o}}{}^2(\hat{o}_{t-1}, a_{t-1})$と表し，4.4.4 項で説明する方法で算出する．$\sigma_a{}^2(\hat{o}_{t-1})$と$\sigma_{\hat{o}}{}^2(\hat{o}_{t-1}, a_{t-1})$ は本手法に合わせた算出方法をとっているためここでは説明を省く．

このとき，認識状態$\hat{o}_{t-1}$を細分化する確率を$\mathrm{P}_{\mathrm{segmente}}(\hat{o}_{t-1})$と表し式(4.2)で算出する．式(4.2)において，$\varsigma_b(x-\theta)$ はゲインb，閾値θのシグモイド関数であり，式(4.3)で表される．シグモイド関数は単調増加の関数であり，x が大きくなるほど 1 に，小さくなるほど 0 に近づく図 4.5 に示すような曲線である．また，変曲点は$(\theta, 0.5)$であり，ゲインbの大小で図 4.5 のように変化する．式(4.3)において，b_N, θ_Nはそれぞれ認識状態の経験回数のシグモイドにおけるゲインと閾値，b_a, θ_aはそれぞれ選択した行動の分散のシグモイドにおけるゲインと閾値，$b_{\hat{o}}, \theta_{\hat{o}}$はそれぞれ遷移した状態の分散のシグモイドのゲインと閾値を表す．これらより，式(4.2)で算出する$\mathrm{P}_{segmente}(\hat{o}_{t-1})$は，全ての経験情報が大きくなるほど 1 に近づく．この場合，選択する行動の分散と，遷移する認識状態の分散が大きいため，認識状態$\hat{o}_{t-1}$は学習に悪影響を及ぼしている可能性が高いと考えられる．一方，閾値より大幅に小さな値をとる経験情報が 1 つでもあれば，$\mathrm{P}_{segmente}(\hat{o}_{t-1})$は 0 に近づく．この場合，小さな値をとっているのが認識回数であれば認識状態$\hat{o}_{t-1}$における経験が十分ではないと考えられ，2 つの分散のどちらかまたは両方であれば，認識状態$\hat{o}_{t-1}$が学習に悪影響を及ぼす可能性は低いと考えられる．以上が細分化を行う確率の決定までの流れであり，図 4.6 にまとめる．

$$\mathrm{P}_{\mathrm{segmente}}(\hat{o}_{t-1}) = \varsigma_{b_N}(\mathrm{N}(\hat{o}_{t-1}) - \theta_N) \cdot \varsigma_{b_a}\left(\sigma_a{}^2(\hat{o}_{t-1}) - \theta_a\right) \cdot \varsigma_{b_{\hat{o}}}\left(\sigma_{\hat{o}}{}^2(\hat{o}_{t-1}, a_{t-1}) - \theta_{\hat{o}}\right) \tag{4.2}$$

$$\varsigma_b(x-\theta) = \frac{1}{1+e^{-b(x-\theta)}} \tag{4.3}$$

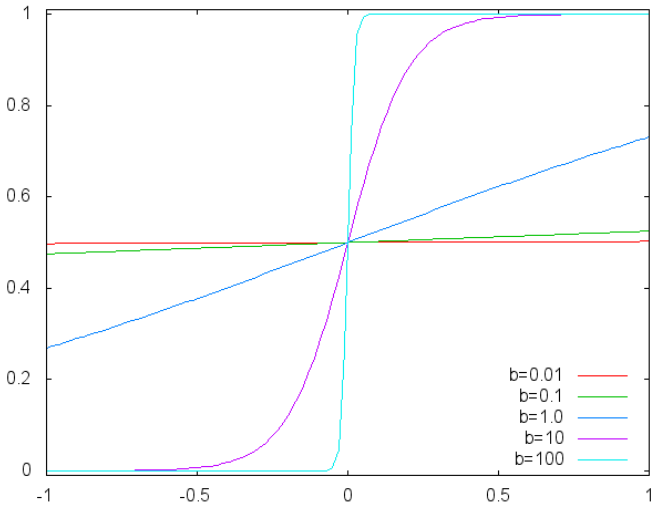

図 4.5　θ =0，b=0.01, 0.1, 1.0, 10, 100 のシグモイド関数

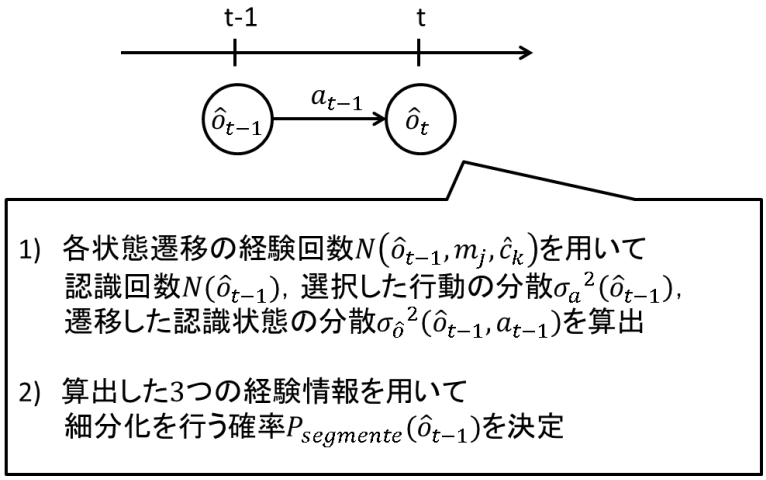

図 4.6　細分化を行う確率を決定するまでの大まかな流れ

4.4.1　経験知識

本研究で用いる経験知識は，ロボットが経験した状態遷移と，過去にその状態遷移を経験した回数を用いて，経験情報蓄積部において知識化する．具体的には，ロボットがある時点tにおいて知識化する経験知識はそれまでの経験知識の数をn_Eとすると，$e_{n_E+1} \in \mathrm{E}$ (Eは経験知識の集合を表す)と表せ，以下の式(4.4)で定義する．

$$e_{n_E+1} = (\ \hat{o}_{t-1}, a_{t-1}, \hat{o}_t\ , \mathrm{N}(\hat{o}_{t-1}, a_{t-1}, \hat{o}_t)\ ) \qquad (4.4)$$

式(4.4)において，$\hat{o}_{t-1}$ は直前の認識状態，a_{t-1}は$\hat{o}_{t-1}$に対してとった行動，$\hat{o}_t$ はその結果遷移した認識状態，$\mathrm{N}(\hat{o}_{t-1}, a_{t-1}, \hat{o}_t)$ は時点tまでにこの状態遷移を経験した回数を表している．つまり，

式(4.4)により定義した経験知識はある状態遷移そのものと，その遷移を経験した回数を表している．ただし，経験知識集合Eに$e_{n_E+1} = (\hat{o}_{t-1}, a_{t-1}, \hat{o}_t, \mathrm{N}(\hat{o}_{t-1}, a_{t-1}, \hat{o}_t))$に一致する経験知識が存在する場合，ロボットは過去にこの状態遷移を経験し，既にこの状態遷移の情報を記憶していることを意味している．この場合にはe_{n_E+1}を新しく知識化せず，一致する経験知識に含まれる遷移回数$\mathrm{N}(\hat{o}_{t-1}, a_{t-1}, \hat{o}_t)$を1回増加して更新する．つまり，遷移回数が増加する度に新しい経験知識として定義することはない．したがって，本研究と先行研究の提案手法をそれぞれ適用したロボットにおいて，経験した状態遷移が同じであれば，知識化する経験知識の数は同数になる．

4.4.2 認識回数

認識回数は，細分化の判断を行う認識状態において，ロボットが過去にどの程度この認識状態を経験したかを評価する経験情報である．本研究では，認識回数が多いほどその認識状態を経験しているため，学習に必要な情報が得られていると考える．

具体的にある時点tにおいて，直前の認識状態$\hat{o}_{t-1}$の認識回数$\mathrm{N}(\hat{o}_{t-1})$は4.4節の式(4.1)で算出する．そして，算出した認識回数$\mathrm{N}(\hat{o}_{t-1})$を用いて，認識状態$\hat{o}_{t-1}$を細分化する確率$P_{segmente}(\hat{o}_{t-1})$を式(4.2)で決定する．

4.4.3 選択した行動の分散

選択する行動の分散は，細分化の判断を行う認識状態において過去に選択した行動の情報から，適した行動が複数あるかを評価する経験情報である．複数の行動が同程度の確率で選択されているほど選択する行動の分散は大きくなる．よって本研究では選択する行動の分散が大きいほど学習が困難な認識状態であると考える．また，選択する行動の分散が0であれば，1つの行動のみを選択することが出来ているため，適した行動が学習出来ていると考える．ただし，学習では局所解に陥らないよう，確率的に探索的な行動をとるため，適した行動が学習出来ても選択する行動の分散は0にはならないと考える．

具体的に時点tにおいて直前の認識状態$\hat{o}_{t-1} \in \hat{C}$で選択した行動の分散${\sigma_a}^2(\hat{o}_{t-1})$を経験知識から算出する方法を説明する．まず，直前の時点$\mathrm{t}-1$の認識状態$\hat{o}_{t-1}$で各行動を選択する確率を式(4.5)で算出する．

$$\mathrm{P}(\hat{o}_{t-1}, m_j) = \frac{\sum_{k \in v} N(\hat{o}_{t-1}, m_j, \hat{c}_k)}{\sum_{j \in u} \sum_{k \in v} N(\hat{o}_{t-1}, m_j, \hat{c}_k)} \tag{4.5}$$

式(4.5)において，m_jはロボットがとり得る各行動を，$\hat{c}_k$はその結果遷移し得る現在の認識状態をそれぞれ表す．また，uはロボットが認識状態$\hat{o}_{t-1}$でとり得る全ての行動の番号を，vは認識状態$\hat{o}_{t-1}$で各行動m_jをとった際に遷移し得る全ての認識状態の状態番号を表している．つまり$\mathrm{P}(\hat{o}_{t-1}, m_j)$は認識状態$\hat{o}_{t-1}$で各行動$m_j$をとった回数を，認識状態$\hat{o}_{t-1}$を認識した回数で割ったものであり，認識状態$\hat{o}_{t-1}$で各行動$m_j$を選択する確率を表している．

次に，各行動m_jと式(4.5)を用いて算出した各行動を選択する確率P$(\hat{o}_{t-1}, m_j)$から，時点tまでに認識状態$\hat{o}_{t-1}$に対して選択した行動の分散 ${\sigma_a}^2(\hat{o}_{t-1})$ を式(4.6)で算出する．このとき，一般的には分散は確率変数の分布が期待値からどれだけ散らばっているかを示す．したがって，式(4.6)で算出される${\sigma_a}^2(\hat{o}_{t-1})$の値はその分布によって異なる．つまり，図 4.7 の左図と右図のような場合には算出される分散${\sigma_a}^2(\hat{o}_{t-1})$の値は異なる．しかし，本研究において分散${\sigma_a}^2(\hat{o}_{t-1})$はその値から「適した行動が 1 つか，複数か」を評価するために用いるため，図 4.7 の左図と右図のような場合には同じ分散によって評価したい．そこで，行動m_jを選択する確率が降順になるよう，行動m_jに割り当て直すことで並び替えを行い(図 4.8)，その後，式(4.6)で分散 ${\sigma_a}^2(\hat{o}_{t-1})$ を算出する．以上の，選択した行動の分散を算出する手順を以下にまとめる．

1) 認識状態$\hat{o}_{t-1}$において各行動m_jを選択する確率P$(\hat{o}_{t-1}, m_j)$を算出する
2) P$(\hat{o}_{t-1}, m_j)$が降順になるよう，行動m_jに割り当て直すことで並び替える
3) 式(4.6)で認識状態$\hat{o}_{t-1}$で選択した行動の分散${\sigma_a}^2(\hat{o}_{t-1})$を算出する

$$ {\sigma_a}^2 = \sum_{j \in u} {m_j}^2 \cdot P(\hat{o}_{t-1}, m_j) - \left(\sum_{j \in u} m_j \cdot P(\hat{o}_{t-1}, m_j) \right)^2 \tag{4.6} $$

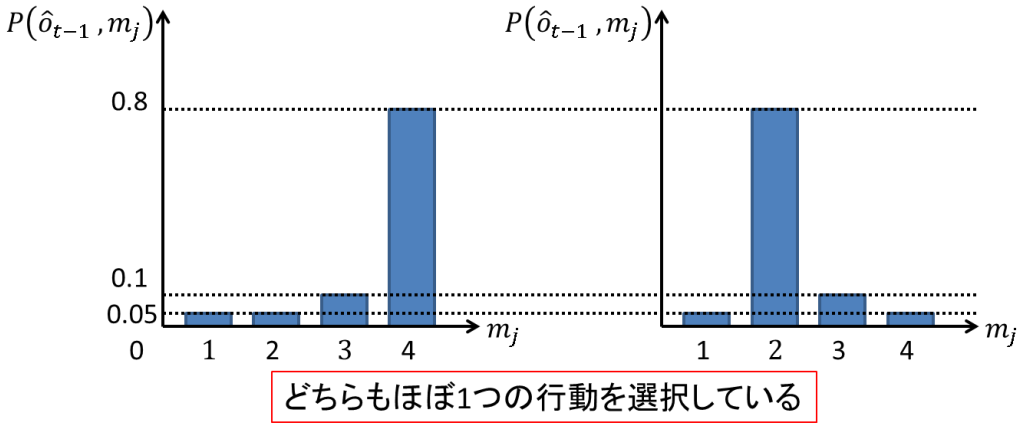

図 4.7　同様に評価したいが分散${\sigma_a}^2$が異なる場合

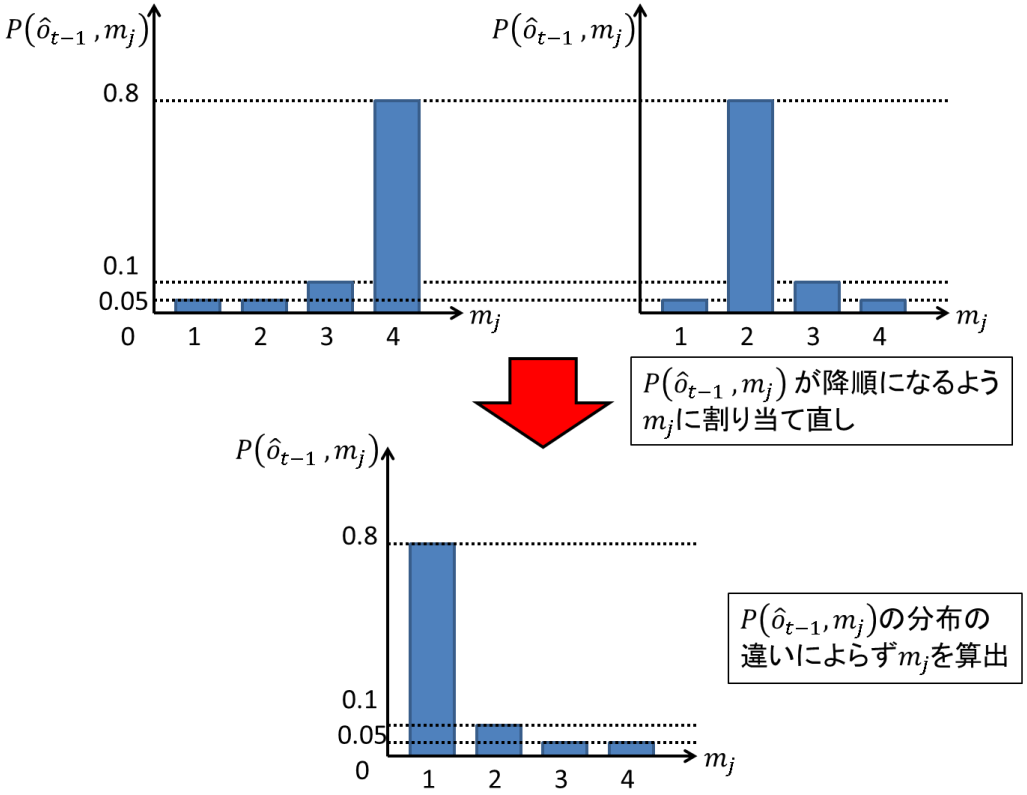

図 4.8　本手法における分散$\sigma_a{}^2$の算出方法

4.4.4　遷移した認識状態の分散

遷移した認識状態の分散は，細分化の判断を行う認識状態において過去に経験した状態遷移の情報から，ある行動をとった際に遷移する認識状態が複数あるかを評価する経験情報である．複数の認識状態に同程度の確率で遷移するほど，遷移した認識状態の分散は大きくなる．よって本研究では，遷移した認識状態の分散が大きいほど，複数の環境状態を混同している可能性が高い認識状態であると考える．また，遷移した認識状態の分散が 0 であれば，その行動をとった際には 1 つの認識状態のみに遷移していると判断できる．さらに，とり得る全ての行動において遷移する認識状態の分散が 0 であれば，その認識状態は不完全知覚による環境状態の混同がない認識状態であると考えることが出来る．

具体的に時点tにおいて，遷移した認識状態の分散を経験知識から算出する方法を説明する．まず，直前の時点$\mathrm{t}-1$における認識状態を$\hat{o}_{t-1} \in \hat{C}$，それに対しとった行動を$a_{t-1} \in M$とする．このとき，各認識状態に遷移する確率を式(4.7)で算出する．

$$\mathrm{P}(\hat{o}_{t-1}, a_{t-1}, \hat{c}_k) = \frac{N(\hat{o}_{t-1}, a_{t-1}, \hat{c}_k)}{\sum_{k \in v} N(\hat{o}_{t-1}, a_{t-1}, \hat{c}_k)} \tag{4.7}$$

式(4.7)において，$\hat{c}_k$は認識状態$\hat{o}_{t-1}$で行動a_{t-1}を選択した結果遷移し得る現在の認識状態を表す．

また，vは認識状態$\hat{o}_{t-1}$で行動a_{t-1}をとった際に遷移し得る全ての認識状態の状態番号を表している．つまり$\mathrm{P}(\hat{o}_{t-1}, a_{t-1}, \hat{c}_k)$は状態遷移$(\hat{o}_{t-1}, a_{t-1}, \hat{c}_k)$を経験した回数を，認識状態$\hat{o}_{t-1}$で行動$a_{t-1}$をとった回数で割ったものであり，認識状態$\hat{o}_{t-1}$で行動$a_{t-1}$を選択した結果，各認識状態$\hat{c}_k$に遷移する確率を表している．

次に，各認識状態$\hat{c}_k$と式(4.7)を用いて算出した$\mathrm{P}(\hat{o}_{t-1}, a_{t-1}, \hat{c}_k)$から，時点 t までに認識状態 $\hat{o}_{t-1}$に対して行動a_{t-1}を選択した結果，遷移した認識状態の分散 ${\sigma_{\hat{o}}}^2(\hat{o}_{t-1}, a_{t-1})$ を式(4.9)で算出する．このとき，分散${\sigma_a}^2(\hat{o}_{t-1})$の算出と同様の理由から，認識状態$\hat{c}_k$に遷移する確率が降順になるように並び替える．さらに，本研究の提案手法において， 状態知識の増加により認識出来る状態の数が増加する．このため，認識出来る状態の数が分散の算出に影響してしまう(図 4.9)．この分散$\sigma_{\hat{o}}^2(\hat{o}_{t-1}, a_{t-1})$の変化を抑えるため，本研究では${\sigma_{\hat{o}}}^2(\hat{o}_{t-1}, a_{t-1})$の算出時に，認識状態の状態番号が小さい順に，一定数の認識状態とその認識状態に遷移する確率のみを用いて分散${\sigma_{\hat{o}}}^2$を算出する．つまり，分散${\sigma_{\hat{o}}}^2$の算出に用いる最大の認識状態の番号をc_{limit}として式(4.8)で正規化し(図 4.10)，遷移した認識状態の分散${\sigma_{\hat{o}}}^2(\hat{o}_{t-1}, a_{t-1})$を式(4.9)で決定する．

$$\hat{c}_k = \frac{k}{c_{limit} - 1} \quad (k = 0,1,\cdots, c_{limit}) \tag{4.8}$$

$$\sigma_{\hat{o}}^{\,2}(\hat{o}_{t-1}, a_{t-1}) = \sum_{k=0}^{c_{limit}} \hat{c}_k^{\,2} \cdot \mathrm{P}(\hat{o}_{t-1}, a_{t-1}) - \left(\sum_{k=0}^{c_{limit}} \hat{c}_k \cdot \mathrm{P}(\hat{o}_{t-1}, a_{t-1}) \right)^2 \tag{4.9}$$

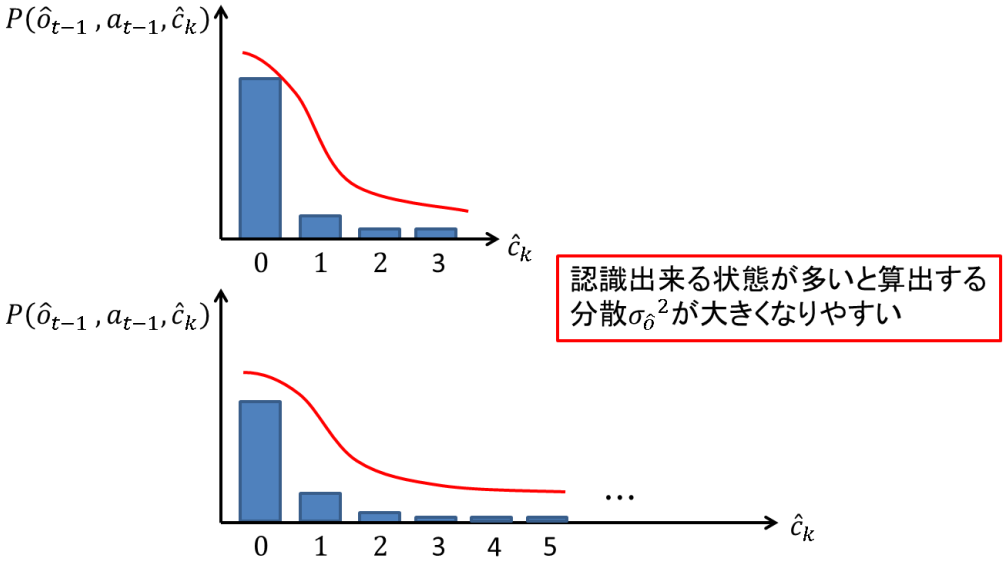

図 4.9　認識出来る状態の増加による分散${\sigma_{\hat{o}}}^2$の変化

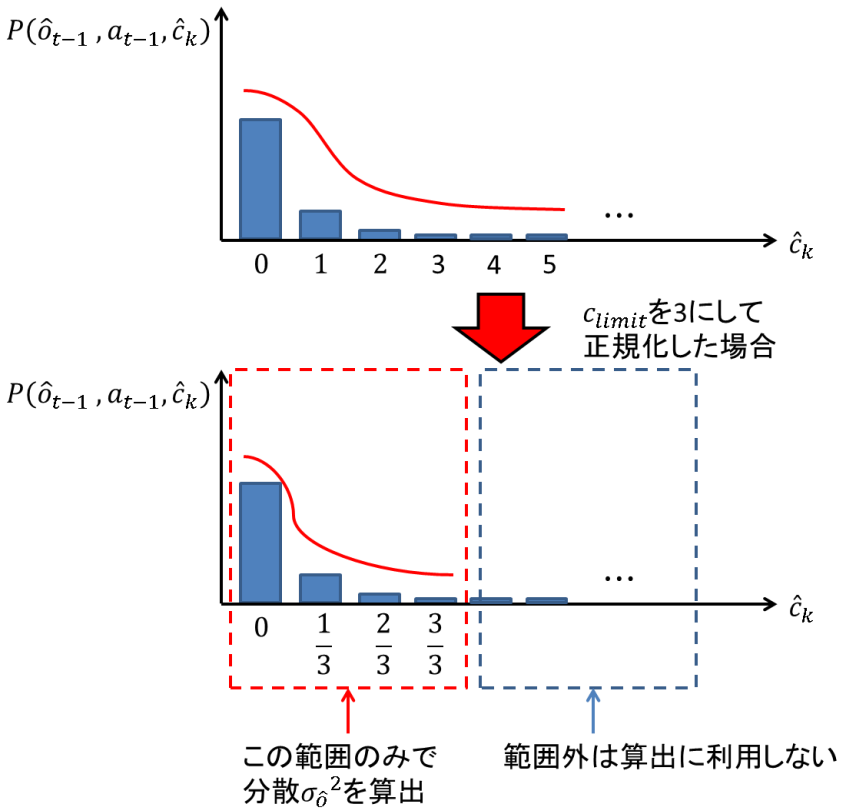

図 4.10　本手法における分散$\sigma_{\hat{o}}{}^2$の算出方法

4.5　細分化判定部

細分化判定部ではある時点tにおいて，その直前の認識状態$\hat{o}_{t-1}$を細分化する確率$P_{segmente}(\hat{o}_{t-1})$を決定し，決定した確率に基づいて認識状態$\hat{o}_{t-1}$を細分化するか判断する．この確率の決定方法については 4.4 節で述べた通りである．このように確率的に細分化を行うことで細分化を抑制し，状態知識の数を抑制することが出来ると考える．

4.6　経験情報蓄積部

経験情報蓄積部では，ある時点tにおいて経験した状態遷移$(\hat{o}_{t-1}, a_{t-1}, \hat{o}_t)$とその状態遷移の経験回数$\mathrm{N}\left(\hat{o}_{t-1}, a_{t-1}, \hat{o}_t, N(\hat{o}_{t-1}, a_{t-1}, \hat{o}_t)\right)$を用いて，式(4.10)で新たな経験知識$e_{n_E+1}$を定義し，4.4.1 節で説明した手順で経験知識の知識化または更新を行う．ただし，n_Eは時点tまでにロボットが記憶した経験知識の数を表す．

$$e_{n_E+1} = (\hat{o}_{t-1}, a_{t-1}, \hat{o}_t, \mathrm{N}\left(\hat{o}_{t-1}, a_{t-1}, \hat{o}_t, N(\hat{o}_{t-1}, a_{t-1}, \hat{o}_t)\right) \tag{4.10}$$

第 5 章　シミュレーション実験

本章では，本研究で提案する手法が知識量を抑制しつつ，不完全知覚に対して有効に働くかどうか，シミュレーション実験を通して検証を行う．実験では，4 章で提案した強化学習に適用したシステムを利用する．実験には比較対象として認識方法の異なる強化学習エージェントを他に 3 体用いる．1 体目は不完全知覚を起こさないエージェント，2 体目は不完全知覚を起こすエージェント，最後に 3 体目は先行研究の手法を適用したエージェントである．これらのエージェントに本論文で提案した手法を適用したエージェントを加え，4 体の学習の様子を比較する．また，先行研究の手法を適用したエージェントと，本研究の提案手法を適用したエージェントの知識量の増加の様子を比較する．

5.1　実験概要

本論文では迷路問題を用いて，提案手法が知識量を抑制しつつ，不完全知覚に対して有効に働くか検証する．シミュレーション実験の概要を図 5.1 に示す．本実験では比較対象を含め，4 体のエージェントを用いる．1 体は不完全知覚を起こさないエージェント，1 体は不完全知覚を起こすエージェント，1 体は先行研究の提案手法を適用したエージェント，最後に本研究の提案手法を適用したエージェントである．これらのエージェントで同一の環境，タスクの迷路問題を行い，その際にエージェントがスタートからゴールに到達するまでにかかった行動数，さらに，先行研究のエージェントと本研究のエージェントの状態知識の増加の様子に注目して比較する．本実験ではエージェントがスタートからゴールに到達するまでを 1 試行とし，各エージェントは各試行終了時，スタート位置に自動的に戻る．

本実験は，不完全知覚が学習に悪影響を及ぼさないタスクと，学習に悪影響を及ぼすタスクの 2 つのタスクを行う．各タスクにおいて環境(迷路)は共通であるが，ゴール位置が異なるために各環境状態(迷路の各マス)でとるべき行動が変化している．これにより，不完全知覚が学習に悪影響を及ぼすかどうか変化する．各タスクにおける実験で，各エージェントの認識方法や学習方法はタスク間で共通であるが，パラメータ等はタスクによって一部変化する．

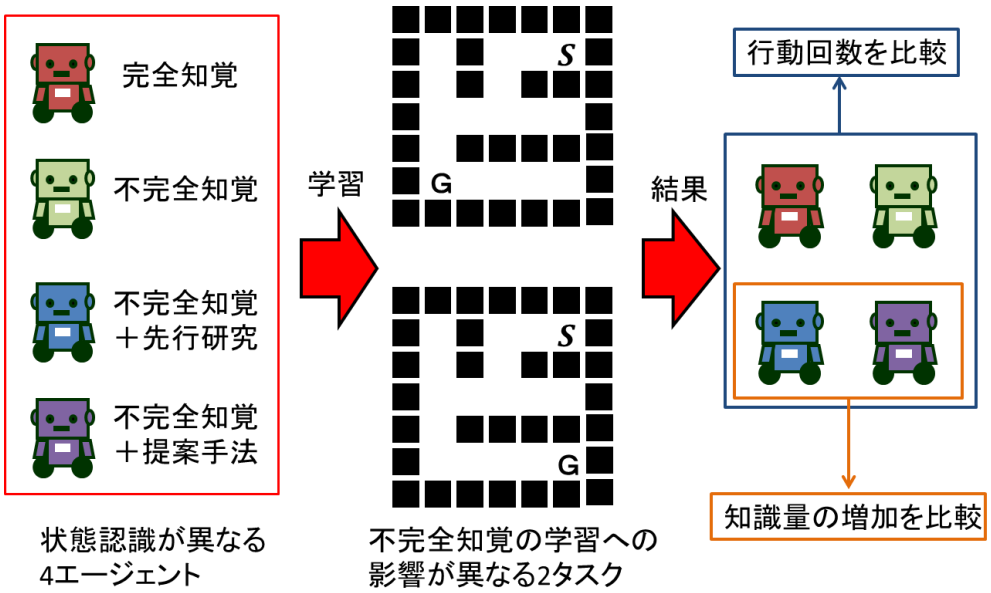

図 5.1　実験概要

5.2　実験目的

本実験により，本研究で提案する手法の有効性を検証する．具体的には，不完全知覚を改善して学習を行うことが出来ているか，先行研究の提案手法に比べて知識量が抑制出来ているかの 2 点に注目し検証する．また，不完全知覚の学習への影響の違いにより，本研究の提案手法の学習の様子や知識量の増加の様子に違いがあるか，注目していく．

5.3　対象エージェント

本実験では以下の 4 体のエージェントを用いる．

・エージェント A：不完全知覚を起こさない
・エージェント B：不完全知覚を起こす可能性がある
・エージェント C：不完全知覚を起こす可能性がある＋先行研究の提案手法
・エージェント D：不完全知覚を起こす可能性がある＋本研究の提案手法

これら 4 体のエージェントは共通して，上下左右のいずれかのマスへの移動という，4 種類の行動を選択することが出来る．ただし，移動先のマスが壁である場合にはその場で待機する．各エージェントの違いは状態の認識方法であり，まず，エージェント A はセンサにより，自身の位置を迷路内の座標として認識することが出来る．これによりエージェント A は迷路内の各環境状態(各マス)を個別に認識することが出来るため，不完全知覚が起きることはない．そして，エージェント B，C，D はセンサ正面の壁の有無を認識することが出来るセンサを 4 つ持ち，それらを用いることで自身の周囲(上下左右のマス)の壁の有無を認識することが出来る．このような状態認識で得られる観測状態は 16 種類である．さらに，エージェント C は壁の有無による状態認識に加え，先行研究の提案手法による状態認識を，エージェント D は本研究の提案手法による状態認識

を行うことが出来る．

5.4　実験環境

本実験で用いた環境を図 5.2 に示す．この環境において，壁の有無によって状態認識を行うエージェントは，同じ色で示したマスで同じ観測状態(■は左右に壁，■は上下に壁，■は上下と右に壁)を得る．そのため，センサによる状態認識では同じ色のマスを区別することは出来ない．この環境でスタート位置やゴール位置を変更することにより，各環境状態で適した行動が変化し，不完全知覚の学習への悪影響の有無が変化する．本実験ではゴール位置を変更することで，不完全知覚が学習に悪影響を及ぼさないと考えられるタスクと，学習に悪影響を及ぼすと考えられるタスクの 2 つのタスクで実験を行う．

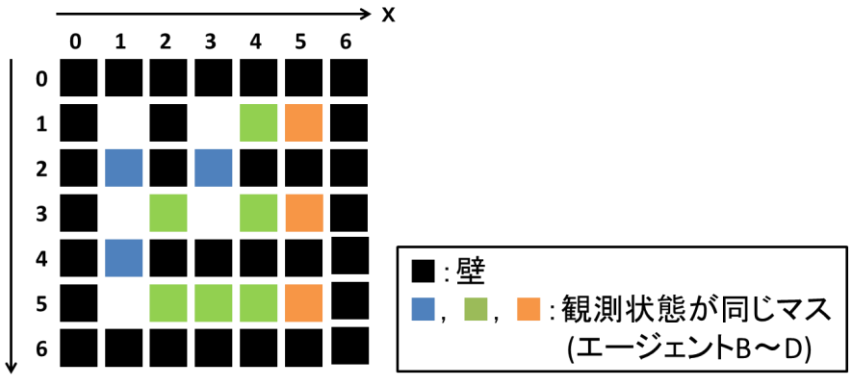

図 5.2　実験で用いた迷路環境

5.5　共通設定

本実験では不完全知覚が学習に悪影響を及ぼさないタスクと，悪影響を及ぼすタスクの 2 つで実験を行う．ここでは，各タスク間で共通に用いる設定について説明する．

5.5.1　エージェント間で共通の設定

本実験では全てのエージェントにおいて，学習に強化学習を用いる．強化学習にはいくつかの学習手法や行動選択手法が存在するが，本実験では共通の学習手法として Q 学習を，行動選択手法として ε -greedy 法を用いる．これらのパラメータは表 5.1 に示した共通のものを用いる．Q 学習では式(5.1)によって学習が進められる．

$$Q(s_{t+1}, a_t) \leftarrow Q(s_{t+1}, a_t) + \alpha \left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_{t+1}, a_t) \right] \qquad (5.1)$$

表 5.1　学習に関する共通設定

試行数	500
学習手法	Q 学習
報酬(ゴール時のみ)	100
α	0.5
γ	0.7
Q 値の初期値	0.001
行動選択手法	ε -greedy
ε	0.05

5.5.2　タスク間で共通の設定

本実験では同じ環境において，ゴール位置の異なる 2 つのタスクで実験を行う．2 つのタスク間で共通する設定として，表 5.2 に示す環境に関する設定と，表 5.3 に示す本研究の提案手法に関する設定がある．

表 5.2　環境に関する共通設定

迷路の大きさ(外壁含む)	7×7
状態数	31(通路のマスの数)
スタート位置(x, y)	(5, 1)

表 5.3　本研究の提案手法に関する設定

経験回数の閾値θ_N	100
選択する行動の分散の閾値θ_a	1.0
遷移する認識の分散の閾値θ_o	0.04
経験回数のシグモイドのゲインb_N	0.3
選択する行動の分散のシグモイドのゲインb_a	30
遷移する認識状態の分散のシグモイドのゲイン$b_{\hat{o}}$	750

5.6　不完全知覚が学習に悪影響を及ぼさない場合：タスク 1

今回，本研究の提案手法の有効性を検証するため，不完全知覚が学習に悪影響を及ぼさないタスクと，不完全知覚が学習に悪影響を及ぼさないタスクの 2 つのタスクで実験を行った．本節ではまず，不完全知覚が学習に悪影響を及ぼさないと考えられるタスクである，タスク 1 について説明する．そして，実験結果として行動回数，状態知識の増加の様子を示し，結果の考察を行う．

5.6.1　タスク 1 で固有の設定

まず，タスク 1 で用いた迷路を，スタート位置とゴール位置と合わせて図 5.3 に示す．図 5.3 において，左図は迷路とスタート位置，ゴール位置を示している．また，右図はスタート位置からゴール位置に到達するために，各マスにおいてとるべき行動を矢印で表している．これに関して，矢印がないマスはゴールを通らずに到達することはないため，実験設定からエージェントが通ることのないマスである．この迷路において，左上のマスを原点として右方向へ x，下方向へ y 座標を持つ．

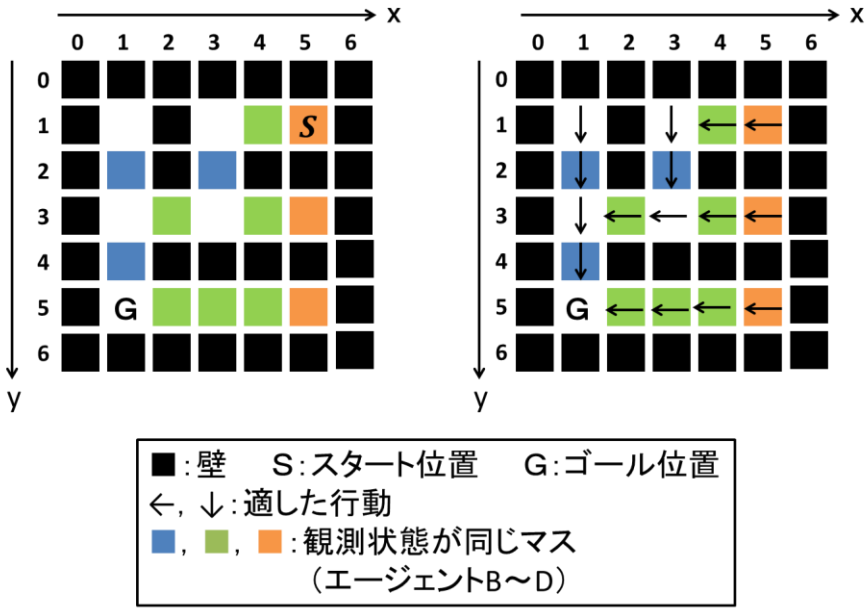

図 5.3　タスク 1 の実験環境

図 5.3 に示した実験環境において，エージェント B～D が不完全知覚を起こすマスは■，■，■で示した 3 つである．これらのマスでを認識した状態において，とるべき行動は図 5.4 右図に示した通り，それぞれ 1 つである（■と■では左へ移動，■では下へ移動）．したがって，タスク 1 の実験環境において，不完全知覚が学習に悪影響を及ぼすことはないと考えられる．タスク 1 で固有のパラメータについて，環境のパラメータを表 5.4 に示す．

表 5.4　タスク 1 における環境のパラメータ

ゴール位置(x, y)	(1, 5)
最短行動数	8

5.6.2　タスク 1 の実験結果

以上の設定で行った実験の結果としてまず，各エージェントの各試行の行動回数を示していく．

各エージェントの各試行における行動回数の推移の比較を図 5.4 に，図 5.4 を行動回数 50 回までの範囲に拡大したものを図 5.5 にそれぞれ示す．また，各エージェントの総行動回数の推移の比較を図 5.6 に示す．次に，エージェント C とエージェント D について，各エージェントの各試行における状態知識の数の推移の比較を図 5.7 に示す．

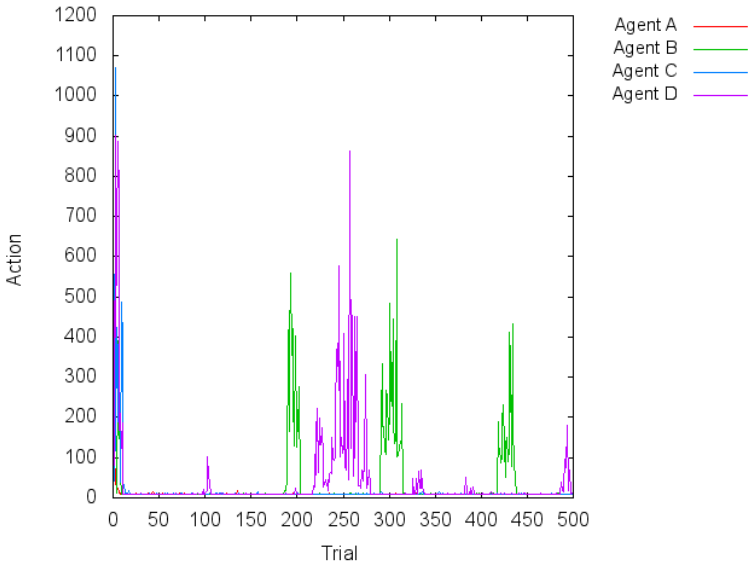

図 5.4　各エージェントの各試行における行動回数の推移の比較

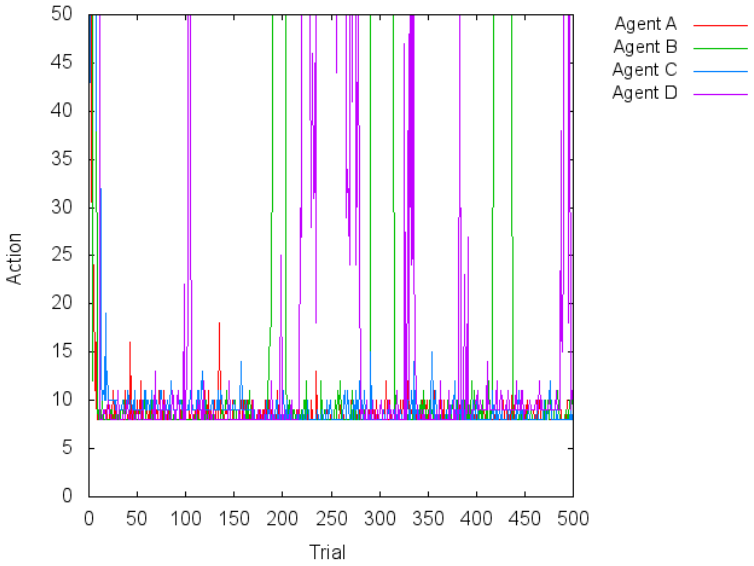

図 5.5　図 5.4 を行動回数 50 回までの範囲で拡大

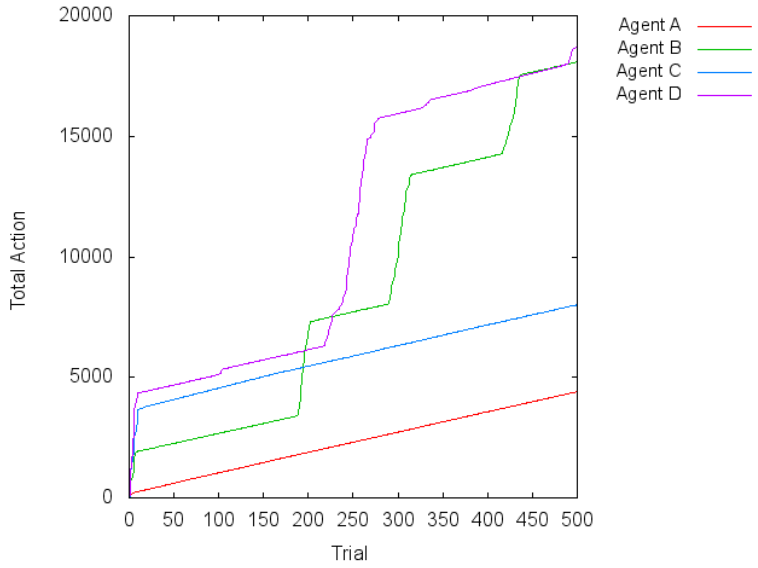

図 5.6　各エージェントの総行動回数の推移の比較

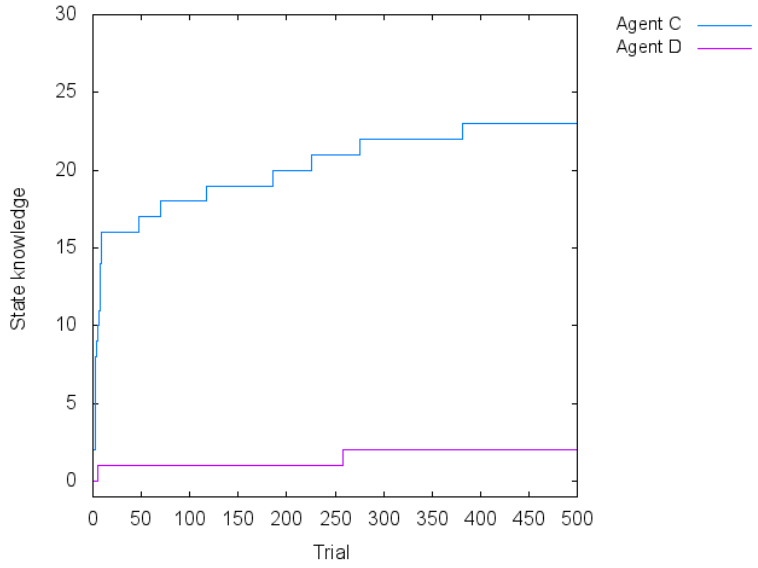

図 5.7　エージェント C とエージェント D の各試行における状態知識の数の推移の比較

5.6.3　タスク 1 の実験結果考察

まず学習結果について，行動回数の推移を示した図 5.4，図 5.5，図 5.6 に注目する．まず，学習の収束について図 5.4 から，全てのエージェントが 20 試行程度で学習が収束していることがわかる．また，エージェント A とエージェント C は学習収束後も行動回数が大きく増えることはなく，エージェント B とエージェント D は学習収束後にも大きく行動回数が増加することがあったことがわかる．この行動回数の増加はエージェントが探索行動により，環境内の学習が十分でない部分に入り込み，試行錯誤を繰り返して学習を行っているためだと考えられる．これは，行動

数の増加が一時的なものであることから判断出来る．また，各エージェントが学習した行動について図 5.5 から，各エージェントの行動回数は 10 行動弱に収束していることがわかる．このことから，各エージェントは適した行動を学習することが出来たと考える．そして図 5.6 から，総行動回数が少なかったエージェントから順に，エージェント A，エージェント C，エージェント B，エージェント D となっており，エージェント B とエージェント D の総行動数はほぼ変わらないことがわかる．この結果は，このタスクでエージェント D がほとんど細分化を行わず，状態知識を持たないため，実質不完全知覚エージェントと認識能力がほぼ変わらないためであると考える．実際に，図 5.7 から，先行研究と本研究の状態知識の増加の様子を比較し考察する．図 5.7 から，エージェント C はおよそ 23 個，エージェント D はおよそ 3 個の知識を記憶していることがわかる．このことから，本研究は先行研究と比較し，学習への影響が無い場合に不要な細分化を抑制することが出来ていると判断する．そして，エージェント C はタスク開始から 20 試行程で状態知識がおよそ 15 程度まで短い試行数の間に増加していることがわかる．これに対し，エージェント D は学習初期と，探索行動を行っていたと考えられる試行数 250 試行程の部分でのみ，状態知識がわずかに増加している．このことから，タスク 1 のような不完全知覚が学習に悪影響を及ぼさないと考えられるタスクでは，本研究の提案手法により，先行研究と比較して状態知識の増加を抑制することが出来ていると判断する．

5.7　不完全知覚が学習に悪影響を及ぼす場合：タスク 2

5.7.1　タスク 2 で固有の設定

まずタスク 2 で用いた迷路を，スタート位置とゴール位置と合わせて図 5.8 に示す．図 5.8 において，左図は迷路とスタート位置，ゴール位置を示している．また，右図はスタート位置からゴール位置に到達するために，各マスにおいてとるべき行動を矢印で表している．

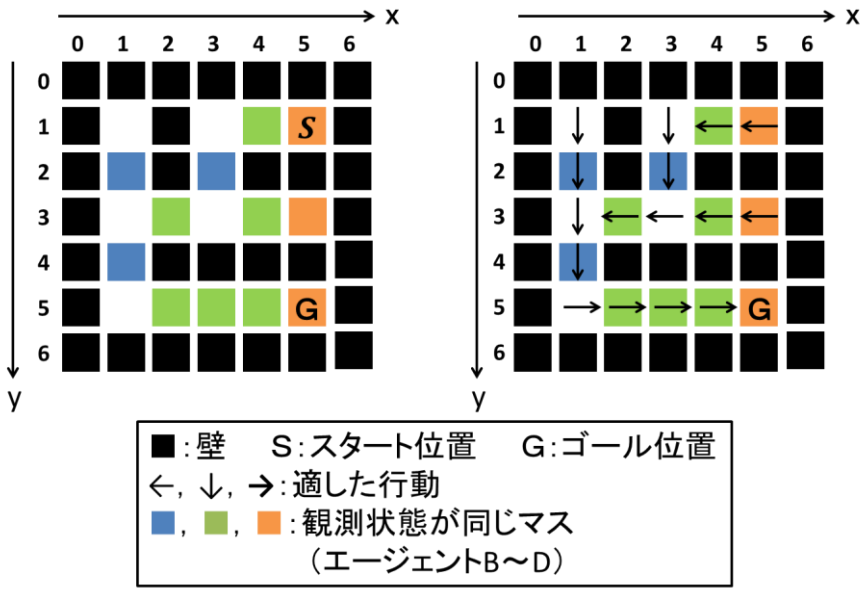

図 5.8　タスク 2 の実験環境

図 5.8 に示した実験環境において，エージェントが不完全知覚を起こすマスはタスク 1 と同様■，■，■で示した 3 つである．これらのマスにおいてとるべき行動は図 5.8 右図に示した通りであり，■と■ではとるべき行動は 1 つ，■では右への移動と左への移動の 2 つである．したがってタスク 2 の実験環境において，不完全知覚が学習に悪影響を及ぼし，■において適した行動が学習出来ないと考えられる．タスク 2 に固有のパラメータについて，環境のパラメータを表 5.5 に示す．

表 5.5　タスク 2 における環境のパラメータ

ゴール位置(x, y)	(5, 5)
最短行動数	12

5.7.2　タスク 2 の実験結果

以上の設定で行った実験の結果としてまず，各エージェントの行動回数の推移を示していく．各エージェントの各試行における行動回数の推移の比較を図 5.9 に，図 5.9 を行動回数 2000 回までの範囲で拡大したものを図 5.10 に，図 5.9 を行動回数 50 回までの範囲で拡大したものを図 5.10 にそれぞれ示す．さらに，各エージェントの総行動回数の推移を図 5.11 に示す．次に，先行研究と本研究エージェントの，各試行における状態知識の数の推移を図 5.12 に示す．

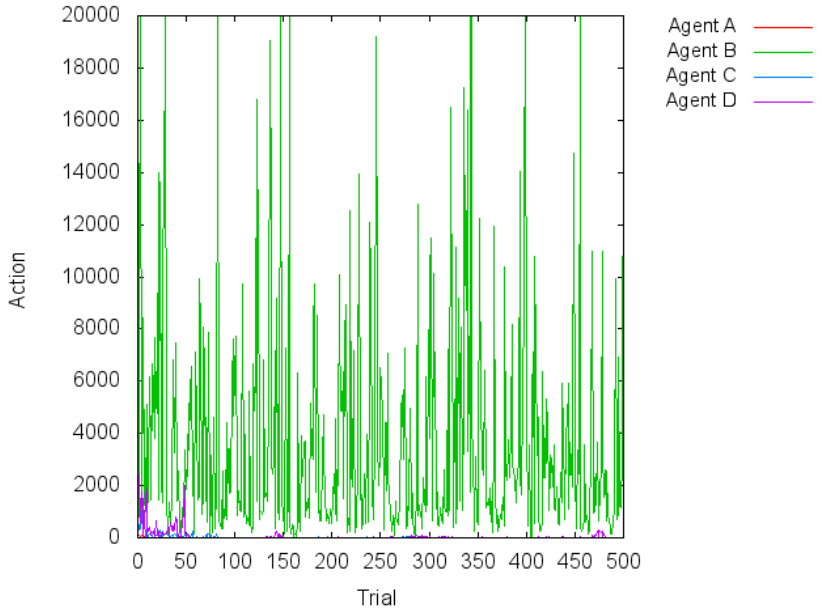

図 5.9　各エージェントの各試行における行動回数の推移の比較

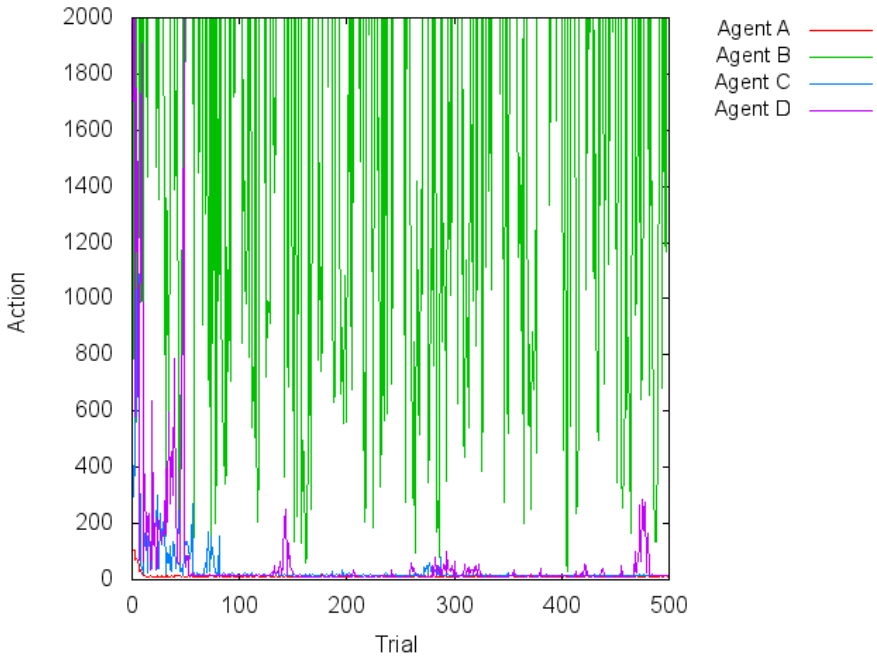

図 5.10　図 5.9 を行動回数 2000 回までの範囲で拡大

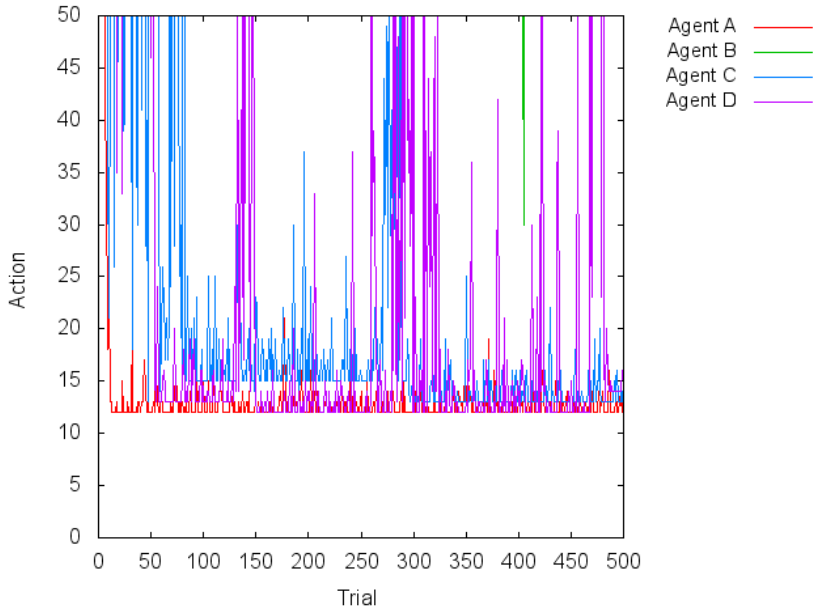

図 5.11　図 5.9 を行動回数 50 回までの範囲で拡大

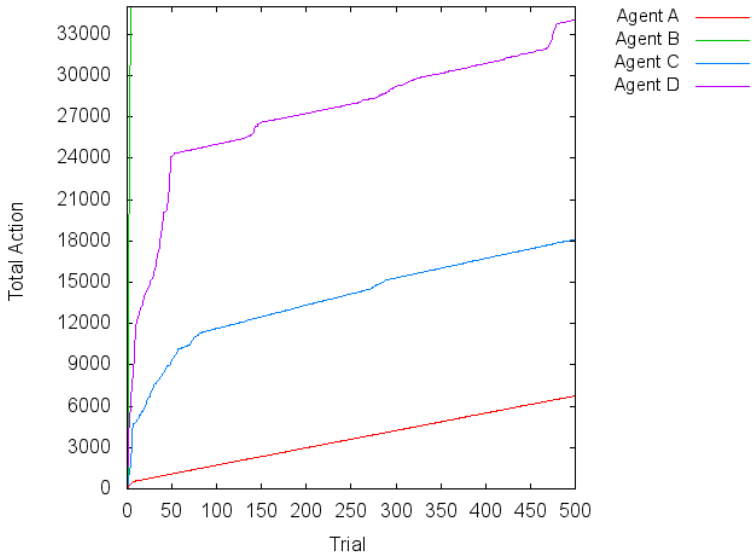

図 5.12　各エージェントの総行動回数の推移の比較

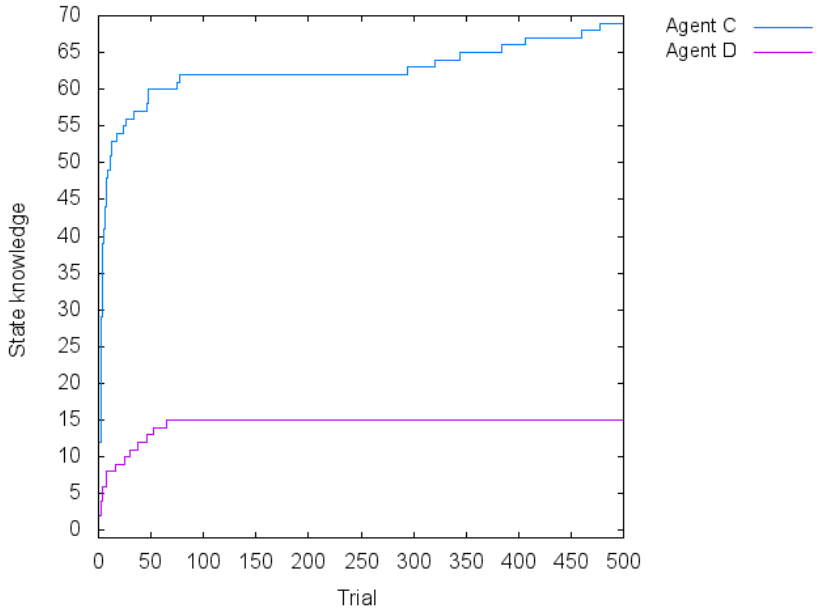

図 5.13　先行研究及び提案手法の各試行における状態知識の数の推移の比較

5.7.3　タスク 2 の実験結果の考察

まず学習結果について，各試行における行動回数の推移を示した図 5.9，図 5.10，図 5.11 に注目する．まず図 5.9 から，エージェント B は行動回数が収束していないことがわかり，不完全知覚により適した行動が学習出来ていないと判断出来る．以降，他のエージェントについて比較，考察を行う．

まず，図 5.10 から，他のエージェントは行動回数が収束していることがわかる．このことから，エージェント C とエージェント D は不完全知覚を改善し，学習が出来るようになったと判断する．さらに図 5.11 に注目すると，まずエージェント A は 20 試行程で学習が収束し，その値は最短行動回数である 12 回程度であることがわかる．次にエージェント C は 50 試行程でおおよそ学習が収束しており，行動回数は同じ試行のエージェント A，エージェント D よりはやや多い 15 回程度であることがわかる．最後に，エージェント D はエージェント C と同様に 50 試行程でおおよそ学習が収束しており，その値はエージェント A とほぼ同じであることがわかる．これらから，エージェント C とエージェント D は不完全知覚を改善する必要がある分，完全知覚よりも多く各環境状態を経験する必要があると考えられる．また，総行動回数について図 5.12 から，各エージェントの実験終了までにとった行動回数は，エージェント A は 7000 回程度，エージェント C は 18000 回程度，エージェント D は 35000 回程度である．さらに，エージェント C とエージェント D について，学習初期である 50 試行程度までに総行動回数が急激に増えており，エージェント C が 11000 回程度，エージェント D が 24000 回程度まで増加している．その後はほぼエージェント A と同傾向で総行動数が増加していることから，エージェント C，エージェント D ともに，不完全知覚を改善するためには多くの経験を必要とすることがわかる．それまでの総行動回数の差は先行研究が確定的に細分化を行っているのに対し，本研究の提案手法では確率的に細分化を行っているた

めであると考えられる．これは実際に，図 5.13 でエージェント C は多量の状態知識を急激に作成しているのに対し，エージェント D は緩やかに状態知識が増加していることからも判断出来る．

第 6 章　結論

6.1　まとめ

本研究では，はじめに，センサを用いたロボットの状態認識には不完全知覚という問題があることを述べ，学習に及ぼす影響から不完全知覚を解決する必要性を述べた．そして，不完全知覚の解決を目指した研究として先行研究を挙げた．先行研究はセンサ情報に加え，ロボットの経験を状態認識に利用する手法を提案した．先行研究は提案した手法により不完全知覚を改善することが出来たが，その一方で，知識量の増加という問題が存在した．この知識量の増加という問題に対し，本研究では先行研究が細分化を確定的に行っていたことに注目し，ロボットの経験情報に基づいて確率的に細分化を行うことをアプローチとして述べた．さらに，アプローチに基づき先行研究を改善することで，知識量を抑制しつつ不完全知覚を改善する手法を提案した．本研究では提案した手法の有効性を検証するため，シミュレーション実験を行った．実験は強化学習を用いて迷路問題を行うものであり，状態認識が異なる 4 体のエージェントを用いて不完全知覚の影響の異なる 2 つのタスクを行い，エージェント間，タスク間で結果を比較することで行った．この実験により，不完全知覚が学習に悪影響を及ぼさないタスクではほとんど状態知識を作らず，不完全知覚が学習に影響を及ぼす場合においては，先行研究と比較して緩やかに，必要な知識のみを作成しつつ，不完全知覚を改善し，学習することが出来たことを示した．このことから，この実験により本研究の手法の有効性を示すことが出来たと考える．

6.2　今後の課題

6.2.1　他環境における有効性の検証

今回有効性の検証のために行った実験では，環境は静的な環境であった．本研究で提案した手法は細分化を確率的に行うようにしたことで，動的な環境において，本研究の提案手法の有効性がより確認出来る実験結果が得られる可能性がある．

6.2.2　実ロボットへの適用

この研究の最終的な目的として，実ロボットへの提案手法の適用を考えている．今回提案した手法は，先行研究に存在した知識量の増加という，実ロボットへの適用を考えた際に考えられる問題に対し，静的な環境ではある程度の改善が行えたと考える．しかし，実ロボットへの適用には他に，動的な環境下における検証と対応，連続的な環境・時間への対応など，様々な課題を解決していく必要があると考える．

参考文献

[1] 浅田稔, 野田彰一, 俵積田健, 細田耕, “視覚に基づく強化学習によるロボットの行動獲得”, 日本ロボット学会誌 vol.13 No.1 pp.68～74 , 1995

[2] 港隆史, 浅田稔, “環境の変化に適応する移動ロボットの行動獲得”, 日本ロボット学会誌 vol.18 No.5 pp.706～712 , 2000

[3] 山田誠二, “行為に基づく環境モデリングのための移動ロボットの進化的行動獲得”, 人工知能学会誌 vol.14 No.5 pp.110～118, 1999

[4] Richard S.Sutton , Andrew G.Barto, “Reinforcement learning An Introduction”, 1998
共訳:三上 貞芳, 皆川 雅章, ”強化学習”, 2001 年 8 月 10 日,第 1 版第 2 版発行

[5] 木村元, 宮崎和光, 小林重信, “強化学習システムの設計指針”, 計測と制御 vol.38 No.10, 1999

[6] 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

[7] 宮崎和光, 小林重信, “Profit Sharing の不完全知覚環境下への拡張：PS-r*の提案と評価”, 人工知能学会論文誌 vol.18 No.5 pp.286-296, 2003

[8] 斎藤宗孝, “不完全知覚問題に対する内部メモリを用いた強化学習法に関する研究”, 情報処理学会研究報告.GI[ゲーム情報学]2004(28), pp.81-87, 2004

[9] 山村忠義, 馬野元秀, 瀬田和久, “段階的な視覚を持つエージェントにおける強化学習について-追跡問題を例にして-”,日本知能情報ファジィ学会誌 Vol.18 No.4 pp.561-570, 2006
情報処理学会研究報告.GI[ゲーム情報学]2004(28), pp.81-87, 2004

[10] 宮崎愛央,” 不完全知覚に対する状態認識法 の提案-経験情報に基づく現状態の推定- “, 室蘭工業大学修士研究論文,2012

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，ご助言をいただいた畑中雅彦先生，本田泰先生，佐賀聡人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いただいた認知ロボティクス研究室の木島康隆さん，梅津祐介さん，北山直樹さん，澁谷和さん，杉本大志さん，高泉昇太郎さん，三浦丈典さん，木村敏久さん，挾間重直さんに感謝致します．