平成１０年度

修士論文

強化学習に基づく能動的認識の自己組織化モデル

電気通信大学大学院 情報システム学研究科
情報ネットワーク学専攻
ヒューマンインターフェース学講座

9651020　高野 光雄

指導教官

阪口　豊

出澤 正徳

磯野 春雄

平成１１年２月３日提出

目 次

第 1 章

序論

1.1 本研究の目的

本研究の目的は能動的認識の立場から人間の認識について考察し，そのメカニズムを明らかにしようというものである．我々は視覚，聴覚などの感覚器官を通して，外界から多くの情報を摂取している．大量の情報が溢れているのに関わらず，情報に溺れてしまうことなく物を見ることや音を聞くことがきわめて容易に自然に行われる．いったいこのような仕組みはどうのようになっており，またどのようにして出来上がったのだろうか？これを実現するさいに考慮すべきことのひとつは，異種多量の情報を処理するための機能である．異なる感覚を統合する必要がある．複数の感覚情報を統合して対象を認識することを感覚統合と呼ぶ．脳は視覚，聴覚，触覚等の感覚器官で捉えた情報を統合し対象に対する《像》を形成し対象を理解する．また，対象の形状や性質について観察するときには，様々な角度から，対象を眺めて立体形状を視覚的に知り，対象にふれたときの感触から材質の形状を感覚的に知る．対象を叩いた場合の音の印象から聴覚的に材質を知る．このように，様々な感覚行為によって得られた情報を組みあわせて対象を理解する．このように様々な感覚情報をを組み合わせて対象を理解する機能は日常生活で重要な役割をはたしいる．

もうひとつ考えられる機能は，対象に関する情報を獲得する際に感覚受容器が捉えた情報を受動的に処理するのでなく，大量の情報のなから必要な情報を選択する機能である．たとえば，対象を観察する場合には，対象を静的に眺めるだけでなく，良く見える位置に頭を動かしたり，対象の特定の場所に視線を向ける．また，文字を読む際には画像情報を一様に処理するのではなく，今，読んでいる文字に注意を向けている．

このように，感覚統合と能動的認識は脳の情報処理を特徴づけている．人間は自分の意図や目的に応じて必要な情報だけを抽出し対象を理解しており，能動的認識は感覚統合と一体となって人間の情報処理を支えている．

人間が外界を理解するうえで感覚統合と能動的認識が重要であることは古くから指摘されてきた．Husserl は，すべての心理的活動は，必ず何かの対象に向けられている (志向性) と指摘し，意識は志向を次々とむけかえることによって対象の全体像を構成すると主張している [2]．心理学者 Piaget は，感覚の発達は運動によってもたらされるものであり，動作能力の発達にしたがって感覚機構は作られていくと述べている [1]．

このように認識過程における能動性の重要性は思想の分野において強く指摘されてきたにもかかわらず，そのメカニズムについては長い間，ブラックボックスとして放置されてきた．これらのメカニズムのを明らかにするために，情報処理過程を思想としてでなく客観的な理論やアルゴリズムとして記述することが不可欠である．そのための方法として計算機的方法は有益な手段である．本研究では人間と同様の振舞いをするシステムを構成することを通じて，そのメカニズムやアルゴリズムを情報処理レベルで明らかにするという方法を用いる．

脳の情報処理メカニズムを取り扱う計算的方法として人工知能と神経回路モデルが挙げられる．人工知能における問題があり，神経回路モデルにおける問題がある．人工知能と神経回路モデルを結ぶ役割として注目されている強化学習がある．強化学習は行動学習に用いられることの多い学習の枠組であるが，近年，強化学習を用いた認識の研究が盛んになってきており [?]，本研究では強化学習を土台にして能動的認識のモデルを構成する．

本研究の目的は能動的認識の立場から人間の認識について考察し，そのメカニズムを明らかしようというものである．本論文では抽象的なアルゴリズムを実現し，メカニズムの本質を抽出することから始める．その第一段階として感覚情報の選択を学習によって自己獲得するモデルを提案する．

モデルの構成には強化学習の枠組を用いた．強化学習とは学習アルゴリズムを指すのではなく問題の設定を指す．目標課題が与えられたとき，目標に向かっていかにして制御されるかを定式化する手段である．強化学習は試行錯誤学習であり，結果が良かったか悪かっただけから学習を行い，何が正解かを教わる教師あり学習とは異なる．強化学習で

は，モデルのインプリメンテーションの方法などに関係なく，制御にかかわる問題だけを扱うように抽象化されている．制御の法則を示している．環境との対話学習を基調としている (図??)．強化学習は近年脳のモデルとしても興味深い仮説が提案されている．

本研究では強化学習により能動的認識を定式化し，具体的なアルゴリズムには TD 学習を用いた．計算機実験による簡単な例題を通してモデルを検討した．今後モデルを発展させるうえでの考察をした．

1.1.1 本論文の構成

本論文は，のような構成になっている．まず第 1 章で本研究の背景について述べる．まず，研究対象である人間の認識について述べ，能動的認識の定式化を行う土台となる強化学習について述べる．第 2 章では，能動的認識のモデルを提案する．提案モデルを簡単な計算機実験を通じて検討する．今後，モデルの拡張を踏まえて実験を通して考察を行う．第 3 章では，実験から得られた結果を踏まえ，一般的考察を行い，第 4 章を結論とする．

1.2 人間の認識

人間は感覚器を通じて外界や対象の情報を獲得し自分の頭の中にその内部イメージと呼ぶべきものを構成する．本論文ではこの内部イメージに対応するものを《内部像》と呼び人間の知覚過程を対象の内部像を構成する過程と定める．内部像は人間の知覚の様々な面をあらわしている．内部像の状態は時々刻々と変化する．たとえば，ある状態はりんごを表している．別の状態はみかんを知覚した状態である．また，別の状態は，みかんともりんごとも知覚できる状態になっている．このように内部像は明確な対象だけを表現したものでなく，明確でない対象も表現している．現実には'みかん'かつ'りんご'であるものは存在しない．内部像は，現実の対象を写し取ったものではなく，外界からの情報を観察者による内部変換を行なった後の解釈を表している．

感覚器の選択は，知覚状態として明確になるまで繰り返し行なわれる．現在のあいまいな内部像から明確な内部像への過程が認識過程である．人間は複数の感覚情報を組み合わせて対象の内部像を構成しておりこれを感覚統合と呼んでいる [3]．

感覚統合の過程は 2 種類に分類できる 1.1．一つは複数の感覚器の情報が並列に獲得されて同時に処理される．具体的には網膜上の細胞からの並列情報の獲得．視覚，聴覚の融合による対象位置の確定などがある．もうひとつは複数の観測行為によって逐次的に得られた情報を組み合わせて処理するものである．具体的には，視覚によって周囲の状況を理解する場合は，頭を動かしたり，身体を移動させたりと生物個体の制約下における感覚器の利用を考えることである．このように考えると以下の三つにわけることができる．

- 空間的な広がりをもつ情報を獲得する．
- 異種感覚を組み合わせた情報を獲得する．
- 生物的制約の下，感覚器の様々な利用形態により情報を獲得する．

さらに人間の情報処理には意識にのぼらない情報処理と意識的なな情報処理がある．視覚を例にとれば，網膜にうつる外界の情報は明暗，色などの特徴が大脳皮質に送られ処理されている．しかしこれを意識することはない．一方視野内の特定物に眼球運動により，網膜上に情報を取り込んだりと，意識的に行われる処理もある．したがって，

- 複数の感覚器を同時に用いた並列的，受動的な処理.
- 複数の感覚器を逐次的に用いた直列的，能動的な処理.

がある.

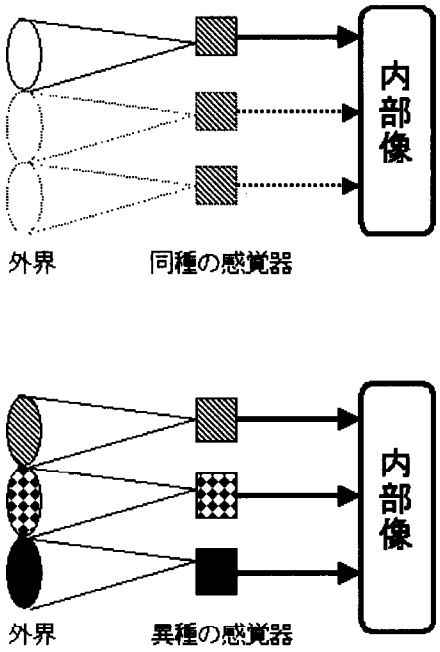

図 1.1: 感覚統合の 2 通りの過程; 実線は感覚器が使用されており，点線は未使用をあらわす.

1.2.1 能動的認識

このように人間の認識を感覚統合という立場で考えた．しかし，多種多様な感覚器の利用による情報の獲得が可能でも，次のような問題が起きると思われる．まず，並列に情報を獲得しても，それらを，すべて同時に処理することが困難であると思われる．また，多様な感覚器利用を行ったとしても，どのような手続きをとるかについては何通りも考えられる．さらに，すべての手続きを行うまで待っていることが現実的なのか，あるいは可能なのかという疑問も残る.

おそらく感覚器が獲得した情報をすべて処理しようとすれば，目的と無関係な情報処理に脳の情報処理に関わる資源の大半を奪われ，立往生してしまうと思われる．したがって感覚器の適宜選択が重要な役割を担っていると考えられる．そのための重要な要素は目的と事象に関わる情報である．たとえば，物を探す場合には，自分の記憶，探す物の性質によって，探す場所を限定することになるだろう．そうした性質と無関係な物を見つけ出す場合は探す場所を限定できずに広範囲にわたって探索することになる．つまりその事象都の関わりの中で，対象についてどれだけの情報をもっているかに関係する．このように目的や状況に応じた感覚器の利用を行うことを能動的認識とよび，感覚統合の構成がスムーズに行われるために，欠かせない仕組みであると考えられる．

能動的認識の実現にさいして問題となることは，感覚器を選択するときに対象や環境に関する情報が十分に得られていないという点である．これらの情報は感覚器を選択し，観察した結果得られるものであるから，このような不十分な情報から感覚器を選択するには，その時点で得られている不十分な情報から推定を行う必要がある．さらに，どの感覚器を選択するかについては，その時点の認識の意図が必要である．人間や動物は生存の目的を達成する過程で，何を認識対象にするかを決めている．人間が目的を達成のために，対象を明らかにし，何をしたらよいのか，またするべきでないのかを判断している．例えば，空腹を食物で満たしたいのか？冷えた身体を暖めたいのか？危険を避けるために，今，いる場所は安全か？などである．この意図にしたがって感覚器の選択は，行われる．心理学では，Gibson のアフォーダンスの概念 [4]，Neisser の知覚循環論 [5] などの考え方がされている．

能動的認識モデルの基盤となるものは環境との対話の中から，認識の意図に沿った感覚器の選択を行うことである．こうした仕組みを人間をはじめとする生物は実現している．いったいどのようにして獲得されたのだろうか？この仕組みを実現するために逐一教わったとは考えにくい．おそらく進化の過程，また生存環境における試行錯誤の結果から，望ましい振舞を獲得したと思われる．環境との対話から学ぶ学習の枠組である強化学習について次節で述べる．本研究ではモデルの構築には，強化学習を基盤にしている．

1.3 強化学習

本章でモデル化の基礎としている強化学習について述べる．まず強化学習の特徴と歴史，実例を述べ，続いて構成要素，定式化，代表的解法，最後に脳モデルとして強化学習について触れる．

環境と対話することによって学ぶという考えは,学習の本質について考えるとき重要な役割を果たすと思われる．実践を通して，行動の結果から，目標を実現するための因果関係に関する情報を生成できる．日常生活を通じて，このような対話が主要な知識の源になっている．私たちのすることに環境がいかにして反応するか，振舞を通じて何が起こるかの影響を探索している．

この章では，相互作用学習についての探求を，人間あるいは動物が学習する方法として，どのような学習状況を想定すべきかをしめすとともに，学習方法のための計算論的アプローチをしめす．

科学上のあるいは経済上の学習問題を解決するための機械モデルの有効性は，数学的解析，あるいは，コンピュータシミュレーションにより評価される．機械学習について，強化学習によるアプローチは，特に相互作用による目標達成学習に焦点をおいている．

強化学習は目標指向の計算的アプローチである．お手本通りの指示，完全な環境のモデル (状態遷移など) に頼らずとも，個体の直接的な環境との対話による学習の強調である．明確な目標の存在と因果関係の察知を内包する状態あるいは行動の価値の考え方が強化学習では中心的な役割を果たす．

1.3.1 強化学習の特徴

強化学習を特徴付けるものは幾通りかあるが，それぞれは密接に関係を持ち，切り離すことはできない．ここでは大まかに4つの特徴を述べる．

強化学習は数学的な報酬信号を最大化するようにして，何をすべきか–状況を行動に写像する–を学習することである．学習者は取るべき行動を教わらないかわりに，多くの報酬を得るための行動を試すことで発見しなければならない．最も興味深い事例とは行動は即座の報酬だけでなく，次の状況だけでなく続いて起こるすべてのものに影響を及ぼすかも知れないということである．これらの2つのきわだった特徴–試行錯誤と遅延報酬–をもつのが強化学習である．

強化学習は，学習アルゴリズムを特徴づけることによって定義されるのではなく，学習問題を特徴づけることによって，定義される．問題を解決するのに適合した，あらゆる良いアルゴリズムを強化学習っであるとみなせる．強化学習についての定式化は本章の後半で述べる．

基本的な考え方は，単純に学習者が直面している実際の問題で，目標を遂げようとする環境との対話の中から，最も重要な面を捕らえることである．そのために学習者はある程度，環境の状態を感知できなければならなく，その状態に影響する行動を取ることができなければならない．また目標達成の経験をしなければならない．定式化は，最も簡単な可能な形で本質を見失わなずに，まさしくこれらの3つ–感覚，行動，及び目標–の面を含むものが意図される．

強化学習は教師あり学習（機械の学習能力におけるほとんどの現在の研究で研究される学習，統計パターン認識及び人工の神経のネットワークの種類）と異なっている．教師あり学習は，ある博識な外部の教師によって提供された例から学ぶことである．これは重要な学習のひとつであるが，環境との相互作用から学ぶことには適切でない．エージェントが活動するすべての状況での正しい振る舞いと，それらの代表的な例を得ることは，相互作用問題については，しばしば非実用的である．相互作用が図式化されていない領域で—エージェントが学習を利益を最大にすることを期待する場合—エージェントは，それ自身の経験から学ぶことができなければならない．

他の学習方には発生せず，強化学習において起こる挑戦課題の一つに探検と開発の間の

tradeoff である. 多くの報酬を得るために, 強化学習エージェントは，かっての試行で見つけた有効な行動を選択しなければならない. しかし, そのような行動を発見するために, それは以前に選択したことがない行動を試みなければならない. エージェントは報酬を得るために，何をするかをすでに知っているが，将来により良い行動選択を行なうためには探検しなえばならない．タスクで失敗せずに開発も探検も追求することはできないという排他的な板ばさみになっている．エージェントは報酬がさまざま試してみて結果がよりよくなるようにする．推計的なタスクでは, 各行動は, 報酬の期待を確実に見積もるために何回も試みられなければならない. 開発と探検のバランスをとる議論は教師学習で起こっていない.

強化学習のもう一つの鍵となる特徴は, 真の目標を明確にして，問題の解決を全体的な立場から扱っていることである．このことは教師あり学習が全体構想を描かずに局所的な課題を考えていることと対照的である．教師あり学習の能力が最終的にどう役立つかを考慮していない．実時間の意思決定の役割，あるいは計画に必要な予測モデルを扱っていない．教師あり学習のアプローチは多くの結果を残したが，副課題に焦点があり制限付きのアプローチである.

1.3.2 歴史

強化学習の歴史には, 主に 2 つの流れがある. ひとつは心理学に端を発する動物の学習の 1 つで, 試行錯誤から学ぶという流れ. もうひとつの流れは, 最適の制御に関する問題の研究である. この問題は環境について十分な知識が与えられたとき最適な行動を計画する方法である. したがって一般に学習を扱っていない. まず, 最適制御の研究が強化学習にとりいられる理由を述べてから, 心理学に端を発する試行錯誤学習について述べる.

最適制御という言葉は 1950 年代, 制御機器を設計するために用いられるようになった. この問題にたいするアプローチ方法が 1950 年代半ばに開発された. この問題を解く方程式は Bellman 方程式と呼ばれる. この方程式を用いて最摘制御課題を解く方法はダイナミックプログラミング [7] として知られている. Bellman は同様に, 最適制御を離散的に, 統計的に扱う場合の最適制御課題をマルコフ決定過程として紹介している [8]. 動的計画法 (dynamic programing) は, 問題が扱う状態の数により指数関数的に計算量が増えるという意味である. Bellman は次元の呪縛 (the curse ofdimensionality) と呼んでいる. しかし, この方法は最適制御課題におけるゆ唯一の方法として広く研究されてきた. 動的計画法方法は繰り返し計算により徐々に正解に達する. これが試行錯誤学習の繰り返し計算と結びつくと考え, 最適な試行錯誤のための計画法としてとりこむ. 強化学習は動的計画法のやりかたを利用するためにマルコフ決定過程として定式化される. 後に [25] によって最適制御課題と試行錯誤学習をつなぐ役割をしている. その後 [24] によって十分ひとつになる.

試行錯誤学習の流れは心理学からはじまる. 試行錯誤学習の本質を述べたのは Edward Thorndike である. ある行動の後によい結果が起これば, 再びその選択を行なう傾向を高めるというものとみなしている. 彼はこれを効果の法則 (Law of Effect) と呼んだ [6].

試行錯誤による学習をコンピュータにプログラミングする考えは Turing らのコンピュータと知性に関する思索に遡る [9]. 試行錯誤学習の計算機による初期の研究で代表的なものは, 1954 年の Minsky の学位論文 [10] と Farley,Clark による論文である [11]. Minsky は論文の中で強化学習による計算モデルを議論し, 彼のモデルをアナログ計算機上に SNARCs と呼んだ構成要素をもとに組み立てた. Farley, Clark は試行錯誤による学習のために設計されたニューラルネットワークマシン上に組み立てた. 強化学習という言葉はこのころエ

ンジニアリング一般に用いられる言葉であった.

このころ Minsky は時間誤差学習 (temporal-difference learning) が重要であることに気付いていた [10]. これは問題解決を行うための一連の行為の中で何が結果を導いたかを推定する問題 (temporal credit assignment) を解く為に予測値の時間的変化の和 (予測誤差) を零にすることが特徴である. 時間誤差学習の起源は心理学の条件付け学習である. Arthur Samuel は 1959 年, Shannon のチェスをコンピュータプログラム化するという提案 [17] に着想をえて時間誤差 の考えを用いた学習システムを提案し, チェッカーゲームプログラムとして実装した [18]. 1961 年 Minsky は強化学習に関連し議論として成功への保証が, 多くの決定の中で, どのように作られ, 配分されるのかという問題を議論している [16].

1955 年, Clark, Farley の関心はパターン認識にむけられている [12]. これは, 強化学習から教師あり学習への移行を意味している. このころ, 多くの研究者達は, 彼ら自身, 実際には, 教師あり学習を研究していたが強化学習を研究していると信じているようだった. たとえば, 人工ニューラルネットワークのパイオニアである Rosenblatt, Widrow, Hoff らは, 強化学習に動機付けられていた. 彼らが用いる用語にも報酬と罰 (rewards and punishments) が用いられるが, 彼らが研究したシステムは教師あり学習でパターン認識や知覚学習に適していた. その結果, 1960～1970 年代に, 純粋な試行錯誤学習はまれだった. Michie は 1963 年に Tic-Tac-Toe ゲームをできるようになるという単純な試行錯誤学習システム (MENACE) を記述した [13]. Widrow, Gupta, Maitra は 1973 年 は強化学習ルールを生成した [15]. これは訓練用事例から学ぶ代わりに, 成功と失敗の信号から学習することができた. この学習形式は選択的自力適応と呼び, 教師を用いた学習の代わりに評価を用いた学習である. 彼らはこのルールをもとに, どうやって blackjack ゲームのプレイを学ぶかを示した. Michie は 1974 年, 人工知能の本質的側面として, 試行錯誤学習の役割を一貫して強調している. [14] 1972 から 1982 年にかけて, Klopf は研究者達が教師学習に注目しているので適応行動の本質的な側面が失われていることを理解していた [26, ?, ?]. 彼によると失われているものは, 環境を悪い状態から良い状態に制御するための, あるいは, 目標に達するための意欲である. これは, 試行錯誤学習の本質的アイディアである. 1981 年に Barto,Sutton は教師学習と強化学習を区別している [20].

1972年にKlopfはTD学習と試行錯誤学習を結びつける試みを行っている[26]. Suttonは動物学習の考え方やKlopfの仕事に強く影響され，Klopfの考えをさらに発展させた．連続した時間における，予測の変化を起因とする学習ルールを記述，動物の学習理論とむすびつけた．Suttonは1981年にactor-criticモデルを提案した[21]．後にこれを洗練させ，古典的条件付けモデルを提案している．神経回路とTD学習の結びつきが研究されている[23]．suttonは1988年にTD λアルゴリズムを開発している[19]. Wittenによる論文にはTD学習規則が含まれている[25].

TD学習と最適制御の流れを結びつけたのは1989年にChris Watkinsの開発したQ-learningである[24]．Watkinsの仕事により機械学習，人工知能，ニューラルネットワークなどで大きな成長があった．1992年Gerry TesauroによるTD-Gammon[22]の成功は，この分野はさらに注目を集めることになった．現在様々な事例が研究されている．交換機のチャンネル割り当て問題[27]，エレベータの制御問題[28]がある．

1.3.3 実例

強化学習を理解する良い方法は，いくつかのサンプルと可能な事例を考えることである．こうした事例が強化学習の発展を導いてきた．

- チェス名人がコマを動かす．その選択は計画–(相手の手や，こちらの手に対応する相手の反応を予想する) と即座の洞察力による判断—(ゲームの本質に関係した配置や動き).
- 石油精製の調節機械，コントローラーは出来た量，コスト，質のトレードオフ関係を最適化する．指定にコストギリギリまで..
- カモシカは生まれて数分で立上り，３０分後にはで走れるようになる．
- ゴミ集めの移動ロボットが新しい部屋にはいるかバッテリ充電のために基地をみつけるか，この決定はどれくらい簡単に，これまでに通った過程で基地を見つけているかに依存.

1.3.4 構成要素

実例にしめすように学習する主体と，これを取り巻く環境がある．強化学習は学習の主体をエージェントと呼ぶ．これはエージェントは，かならずしも人間や動物をさすわけではない．環境を物理的な外界とみなすのではない．エージェントと環境の境界線をロボットや動物そのものと考えるのは正しくない．境界線はロボットや動物自身の内側にある．ロボットのモーターと稼働部分，センサー機器はエージェント部分というより環境に区分する．人間や動物の骨格，筋肉，感覚器官も同様である．なにをエージェントとするかの規則は,エージェントが任意に変えることができないものをエージェントの外において，これを環境とする．我々は,環境における何もかもがエージェントにとって未知であると仮定せずに，少しばかりは知ることができる.例えば，それはエージェントの取る行動と状態の組合せに応じて受け取る報酬である．

報酬関数 は強化学習における目標を定義する．知覚した環境内の状態 (あるいは状態と行動のペア) を数値に対応させる．エージェントにとって魅力的な状態を指している．エージェントの目標は長い目で見た場合に環境から受け取る報酬合計を最大化することである．報酬関数 は操作の結果が良いことか悪いことなのかを定義する．生物学ではでは報酬を喜びや痛みとみなせるだろう．たとえば，ある方針のもとで選択された行動が，より少い報酬しか受け取れなければ，再び，その状況になった場合は他のアクションを選ぶように修正される．報酬は自然界でも工学でも物理的な個体の内部になるだろう．しかしエージェントの外にある．報酬をエージェントの外，環境の中に定義するのは，エージェントの究極の目標が制御不完全なものではずだとという点にある．エージェントは，単純に，同じようにやったからといって報酬を受け取ることはできない．目標の概念を報酬で形式化することは，強化学習で，最も特徴的なことのひとつである．

方針関数はエージェントの各時刻の振る舞いかたを定義する．粗く言うと，方針関数は知覚した環境の状態とそのときの行動との対応付けである．これは心理学で言うところの刺激と応答の条件付けである．方針関数は，ある場合には簡単な関数や参照表 (lookup table) でも良い．

報酬関数は直後の状態の良さを表す．エージェントはある方針にしたがって行動することになるが，価値関数は長い目で見た状態の良さを表す．状態の価値はエージェントが将

来にわたって蓄積することを期待できる報酬の合計である．それゆえに，報酬は直後の環境状態の固有の魅力を決定している．価値は，その状態の後に続くものと同等にみなし，長期的魅力を決定している．例えば，ある状態の直後の報酬が小さくても，高い価値をもっている．なぜなら，別の状態が続き高い報酬が規則的に生じる．これと逆のことも言える (高い報酬のとき未来の価値は低くなり，つまり目標が達成された)．人間を例にこれを類推すると，高い報酬は喜び，低い報酬悲しみである．価値は，人間の生存環境の特定の部分が，どの程度の喜びと悲しみかを判断する，洗練された長期的な判断に相当する．

報酬は一次的で，報酬予想の価値は二次的である．報酬がなければ価値もない．価値を見積もるのは，より多くの報酬を得るためである．しかしながら決定と評価のさいに最も注意するのは価値である．行動選択は価値判断に基づく．行動系列が長い目で見て，最大の報酬合計値を手いにいれるので，最も高い報酬ではなく，最も高い価値をうみだす行動系列が求まる．不幸なことに，報酬を決定するより価値を決定するほうが難しい．報酬は直接，環境から与えられる．しかし価値は見積もらなければならない．エージェントの存続期間中に行なった観察の系列から再び見積もり直さねばならない．行動決定と計画においては，価値から生成される量に最も多くの関心がある．ほとんどの強化学習アルゴリズムはで重要な部分は，効率的に価値を評価する部分である．近年，価値関数が中心的役割を果たすことが最も重要であるとわかってきた．

本節でとりあげている強化学習は価値関数の評価にもとづいて構築されている．強化学習課題を解く為には厳密には，かならずしも，これを必要としない．たとえばGA，SA，その他の最適化法があ強化学習課題の解法に用いられる．これらの方法は直接に，価値関数の助けを借りずに方針空間を探索する．これらの方法を進化論的 (evolutionary) 方法と呼ばれる．なぜなら，これらの手続きは生物の手法に類似している．個体の存続期間中は学習を行なわず，生物的進化により特別な生態 (あるいは行動) をもった組織を作るやりかたである．もし方針空間が十分な小ささであれば良い方針をみつける．進化論的方法は効果がある．付け加えて言うと，進化論的方法は学習エージェントが環境の状態を正確にセンシングできないときに活用できる．

それにもかかわらず，強化学習として扱うものは環境との対話をしながらの学習を含んでおり，進化論的方法はこれを含んでいない．進化論的方法は強化学習課題の持つ多くの

役立つ構造を無視する．エージェントが探し求めた方針が状態から行動への関数である事実を利用しない．これまでの状態の経緯について注目しない．進化と学習は多くの共通点を共有し，自然界ではそうであるように，一緒になって機能しているのであるが，この本節では”強化学習”の言葉をに進化論的方法を含めない．

強化学習の4番目の要素は環境のモデルである．モデルはプランニングに役立てられる．将来に起こりうる状況を考慮した一つのアクションの系列を決定する．初期の強化学習では，明示的に，トライアルアンドエラー学習だった．これはプランニングとはほぼ対極にあるものだ．しだいにDPと関係するようになる．DPはモデルを用いている．モデルとプランニングの合同は比較的最近のことである．

エージェントが環境と対話するためのしくみがひつようで，エージェント (agent) と環境 (environment) をさらに方針関数 (a policy function), 報酬関数 (a reward function), 価値関数 (a value function), 環境モデル (a model of the environment) の4つに区分した．

1.3.5 定式化

離散時間 t と各時刻の状態 s_t, 行動 a_t, 報酬 r_t を定義する．時間経過につれて以下の手続きで信号系列を生成する．

1. 方針に基づき行動 a_t を決める．
2. a_t を実行
3. 状態が s_{t+1} になったことを知り，報酬 r_{t+1} を受け取る．
4. 方針の改良

以上をくりかえす．この様子を図 1.2に示す．強化学習はエージェントと環境が対話により学習を行う仕組みである．

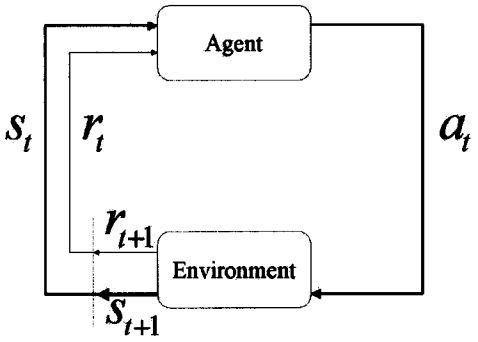

図 1.2: エージェントと環境の対話構造

強化学習の扱う課題は 2 通りある．1つはゲームの勝敗や迷路での目標到達．これらはエージェントと環境の対話の中断が自然に起こる事象である．これをエピソード課題と呼ぶ．もうひとつはロボットの制御などの永続的な性質をもった課題である．これをコンティナス課題と呼ぶ．エージェントと環境の対話の中断が起こらないことが前提となっている．この 2 種類のタスクを扱うために t 以降の報酬を $r_{t+1}, r_{t+2}, ...$ とすれば時刻 t 以降の報酬の蓄積を

$$R_t = \sum_{k=0}^{T} \gamma^k r_{t+k+1}$$

で定義する．これをリターンと呼ぶ．エピソード課題ではTを有限とした場合である．コンティナス課題では$T=\infty$とした場合である．γを割引率と呼ぶ．$0<\gamma\leq 1$である．現在時刻tから遠い将来ほど影響を少くすることで無限大の時間の蓄積を有限量におさめている．

次に価値関数を定義する．エージェントに与えられた状態sの良さは方針πにしたがった場合のリターンの期待値であらわす．また，状態sで行動aの良さは，方針πにしたがった場合のリターンの期待値であらわす．方針πより生成される報酬系列の価値

$$\begin{aligned} V^{\pi}(s_t) &= E_{\pi}(R_t|s_t=s) \\ &= E_{\pi}(\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}|s_t=s) \end{aligned}$$

$$\begin{aligned} Q^{\pi}(s,a) &= E_{\pi}(R_t|s_t=s,a_t=a) \\ &= E_{\pi}(\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}|s_t=s,a_t=a) \end{aligned}$$

価値関数の最大化する方針が最適方針である．最適方針を$*$であらわす．強化学習の目的はこの2つの関数を使うと次のようにあらわせる．

$$V^*(s)=\max_{\pi} V^{\pi}(s)$$

$$Q^*(s,a)=\max_{\pi} Q^{\pi}(s,a)$$

両式の関係は

$$Q^*(s,a)=E\{r_{t+1}+\gamma V^*(s_{t+1})|s_t=s,a_t=a\}$$

状態信号の性質について述べる．状態の性質はマルコフ性を満たすものを扱う．マルコフ性とは時刻tにおいて$t+1$の事象を予測可能な性質をいう．マルコフ性を満たすことで，状態遷移確率$P^a_{ss'}$や報酬分布$R^a_{ss'}$を定めることが可能になる．方針を確率$\pi(s,a)$であらわすと，価値関数は

$$\begin{aligned} V^{\pi}(s) &= E_{\pi}(R_t|s_t=s) \\ &= \sum_a \pi(s,a)\sum_{s'}\wp^a_{ss'}[\Re^a_{ss'}+\gamma V^{\pi}(s')] \\ &= \sum_a \pi(s,a)Q(s,a) \end{aligned} \tag{1.1}$$

であらわせる．最適方針*では次式がなりたつ．

$$\begin{aligned} V^*(s) &= \max_{a \in A(s)} Q^*(s,a) \\ &= \max_a E\{r_t + \gamma V^*(s_{t+1}) | s_t = s, a_t = a\} \\ &= \max_a E\{P^a_{ss'}[R^a_{ss'} + \gamma V^*(s')]\} \qquad (1.2) \\ Q^\pi(s,a) &= E\{r_t + \gamma \max_a' Q^*(s_{t+1}, a') | s_t = s, a_t = a\} \\ &= \sum_{s'} P^a_{ss'}[R^a_{ss'} + \gamma max_{a'} Q^*(s', a')] \qquad (1.3) \end{aligned}$$

これをベルマンの最適方程式という．強化学習は，この方程式を解く問題として定式化される．マルコフ決定過程において最適方針を求めるための解法にはDPがある．DPでは，ベルマンの最適方程式を満たすように次の操作を交互に行う．

- 現在の方針で価値関数を決める．ベルマンの再帰式 (式1.1) を使う．
- 価値関数に基づいて方針を決める．行動価値が最大になるように方針を変更

$$\pi'(x) = \arg\max_a Q^\pi(s,a)$$

式1.1は再帰式である．見積もろうとする状態の価値が別の状態の価値に依存するしくみになっている．したがってすべての状態について繰り返し計算を行うことになるから計算量が増大してしまう．また，完全な環境のダイナミクスの知識を必要しており，学習を扱っていない．

しかし行動計画における繰り返し計算は試行錯誤学習における繰り返しとマッチする．したがって環境の知識が不足している問題の解法の下地になる．DPの理論はマルコフ決定過程における解法としてだけでなく強化学習課題をつらぬく方法でもある．

1.4 強化学習の解法

前節では強化学習を定式化した．マルコフ決定過程の下で環境のダイナミクスの知識が完全である場合，最適な行動を決定するための計画法としてDPがある．本節では，環境のダイナミクスの知識が与えられない場合に，どのようにしたら最適な行動を決定できるかという問題を扱う．

1.4.1 モンテカルロ法

本節で述べるモンテカルロ法はダイナミックプログラミングで用いるしくみをを受け継いでいる．すなわち，方針評価プロセスが価値関数の見積もりを行い，方針改良プロセスは価値が大きくなるように行動選択を改良する．改良された方針を用いて価値関数をふたたび見積もる．方針評価プロセスと方針改良プロセスを交互に繰り返すことである．

方針評価

価値関数の見積もりについて述べる．DPでは環境の知識として，状態遷移確率と報酬存在確率が与えられた．このためDPは状態の価値を見積もれば，その状態で選択可能な行動の価値を将来に受け取る報酬の期待値として知ることができた．行動選択の方針は価値を最大にするように改良した．しかし，環境の知識が与えられなければ状態の価値を見積もって行動の価値を知ることができず，方針を改良することができない．したがって行動と状態が結びついた状態-行動価値を見積もることになる．方針πで状態sで行動aを選択する価値について見積もる手続きについて述べる．

状態sで行動aを選択することを(s,a)の発生と定義する．サンプルリターンを明確に見積もるために学習者の経験をエピソードに分割して扱う．学習者はいかなる行動を選択したとしても最終的にエピソードを終えることができるとする．方針πのもと，状態sにおける行動aの価値を見積もるには，(s,a)の発生を含むいくつかののエピソードを用いる．多ければ多い程正確な価値を見積りが可能になる．あるエピソードにおける(s,a)発生後のリターンは，訪問時刻をtとしてエピソードの終了時刻がt+nならば，その間，観測された報酬から，前節の定義より求まる．これをR_tとする．エピソードがN個あれば

サンプルリターンの平均が (s,a) の価値となる.

$$Q_N(s,a) = \frac{R_1 + R_2 + R_3 + R_N}{N}$$

で求まる. 全てのエピソードについて, リターンを計算するまで待っていては計算コストがエピソードの増加にともなって増加してしまう. 次式を用いて計算する.

$$Q_{k+1}(s,a) = Q_k(s,a) + \frac{(R_{k+1} - Q_k(s,a))}{k+1}$$

k+1 回目のエピソードが終了するたびに見積もることができる. k+1 をαで置き換えて表す. R_{k+1}を Rに置き換える.

$$Q_{new}(s,a) = Q_{old}(s,a) + \alpha(R - Q_{old}(s,a)) \tag{1.4}$$

$\alpha \to 0$ により収束する. モンテカルロ法の特徴は経験から学ぶことができるので環境のダイナミクスについて完全な知識を必要としない. このことは環境についての知識を明確にできない複雑な事例でも扱うことができることを意味する. また, ダイナミックプログラミングのようある状態の見積もりが別の状態の見積もりに依存していない. このことは計算コストが状態の数とは無関係であることを意味する. そして, 扱う事例が大規模な場合は問題を小さな規模に分割してあつかうことが可能である.

次に価値を利用して方針を決定する方法について述べる. エピソード生成は方針に依存する. (s,a) の見積もりのために適切なエピソードを抽出することが必要である. 状態と行動が結びついた価値を見積もっているので最大の価値をもつ行動を選択すれば良い. しかし複雑な問題が生じる.

方針改良

いくつかのエピソードを抽出して, これらを用いて状態-行動価値を見積もる. この中で状態行動価値が最大になっているものがあると思われる. たとえば, 状態 s に関連した行動 a_1, a_2, a_3を仮定し, $Q(s,a_1), Q(s,a_2), Q(s,a_3)$ を見積もった結果, $Q(s,a_1)$ がそれだとする. このとき a_1を状態 s における greedy action と呼ぶ. greedy action を選択することは, これまでの経験の利用 (exploitation) である. nongreedy action を選択することを

探索 (exploration) である．行動選択が，これまでの経験と無関係であるからである．これによって，今より良い結果に結びつく行動を発見できる．

強化学習では，学習者は，これまでの経験を利用し有効な行動を選択する．しかし，そのような行動を発見するには以前の経験とは無関係に行動を試すことになる．greedy action を確定するには他のアクションの評価を十分に行う必要がある．そのためには経験の利用と経験と無関係な探索のバランスをとる **双対問題** が発生する．方針はこの問題に応えるものでなければならない．4 つの解決方法を挙げる．

探索開始の仮定 エピソード抽出の開始においては，行動選択確率は非零にしておく．無限大の試行で全ての行動が試されることを保証する．この方法は探索開始仮定 (exploring starts assumption) である．

1. すべての (s,a),$s \in S, a \in A(s)$ について初期化する．
   - $\pi(s), \pi(s,a)$ を適当に定める．
   - reruns(s,a) を空リストにする．
2. 永遠に繰り返す．
   (a) 探索開始の仮定とπを用いてエピソードを生成する．
   (b) エピソードの中に (s,a) が現れたら
   - R にリターンを入力．
   - R をリスト reruns(s,a) に追加．
   - Q(s,a) にリスト内のリターンの平均を代入

   (c) エピソード内のすべての s について$\pi(s) \leftarrow arg\max_a Q(s,a)$

ただし，見込みのない仮定は環境と直接相互作用する場合には役に立たない．

ε-greedy 法 大部分の行動選択には，greedy に振る舞い，しかし，時々，少ない確率 ε で無作為，一様にこれまでの経験と独立に行動価値を見積もるための行動を選択することである．この方法は，ε-greedy 法である．

1. すべての (s,a),$s \in S, a \in A(s)$ について初期化する.

   - $\pi(s)$ を ε -greedy に定める.
   - reruns(s,a) を空リストにする.
   - Q(s,a) を適当に定める.

2. 永遠に繰り返す.

   (a) πを用いてエピソードを生成する.

   (b) エピソードの中に (s,a) が現れたら

   - R にリターンを入力.
   - R をリスト reruns(s,a) に追加.
   - Q(s,a) にリスト内のリターンの平均を代入

   (c) エピソード内のすべての s について $a^* \leftarrow arg\max_a Q(s,a)$

   - $if(a = a^*), \pi(s,a) \leftarrow 1 - \epsilon + \frac{\epsilon}{|A(s)|}$
   - $if(a \neq a^*), \pi(s,a) \leftarrow \frac{\epsilon}{|A(s)|}$

softmax 行動選択法 ε -greedy 法の欠点は探索のときすべての行動が等確率で選ばれることである. これは, 最も悪いはずの行動が選ばれてしまうことを意味する. 望ましいのは行動選択の価値に応じて選ばれることである. 行動価値の相対的な違いを反映した確率にもとづいて行動が選択される方法は softmax 行動選択法である. greedy action が最も選ばれやすく他のものは価値の大きさに応じて段階的に確率が下る. 最も一般的なものはボルツマン分布によるもので確率を次式で与える.

$$\pi(s,a) = \frac{e^{\frac{Q(s,a)}{\tau}}}{\sum_b e^{\frac{Q(s,a)}{\tau}}}$$

ここで,t 回の行動を試した結果を $Q_t(s,a)$ で表し, τを温度係数と呼ぶ. τは正の数である. τが高ければ全ての行動は等確率で選ばれる. τがある程度低ければ価値の大きさに応じて選ばれ, $\tau \to 0$ ならば greedy になる.

追求法 価値が最大になる行動を選択する傾向を強める．この方法は追求法である．状態 s での行動 a の選択は確率的に決まる．$\pi(s,a)$ であらわした．時刻 t で状態 s において行動 a の価値が最大であれば

$$a^*(t) = arg\max_a Q_t(s,a)$$

であらわす．このとき最大の行動 a^*を選択する確率を次式により高くする．

$$\pi(s,a^*) = \pi(s,a^*) + \beta(1 - \pi(s,a^*))$$

これ以外の行動については，次式により確率を低くする．

$$\pi(s,a) = \pi(s,a) + \beta(0 - \pi(s,a))t$$

1.4.2 TD 法

TD 法はダイナミックプログラミング法に対して有利である．それは，環境のモデルを必要としないからである．モデルとは環境の報酬と，次の状態についての確率分布である．モンテカルロ法がダイナミックプログラミングに対して持っていた有効性を受け継いでいる．

次に，モンテカルロ法に対して有利である．モンテカルロ法がエピソードの終了を待っていなければならないのに対して，TD 法は 1 時刻だけ待てばよいことにある．ある事例のエピソードの中には，とても長いものもあると思われる．エピソードが終わるまで学習が起こらないから収束が遅くなる．また別の事例では，それが continus タスクの場合，エピソードは全く持たない．結局，モンテカルロ法はエピソードの中断を考慮しなければならない．このことは著しく学習を面倒にするだろう．TD 法はブートストラップの性質をもち，各状態遷移後に何が起こるがという点を考慮するから，このような問題の影響を受けにくい．

方針評価

方針πが最適であれば，ベルマンの最適方程式より

$$Q_\pi(s_t,a_t) = r_{t+1} + \gamma Q(s_{t+1},a_{t+1})$$

が成り立つ．モンテカルロ法の方針評価はリターンが発生するまで行なわない．エピソード終了時刻を$t+n$とすれば，その前の時刻$t+n-1$とについて

$$Q(s_{t+n-1}, a_{t+n-1}) = r_{t+n} + \gamma Q(s_{t+n}, a_{t+n})$$

が成り立つ．エピソード終了時点では，これ以上未来に報酬の見込みをがないから，

$$Q(s_{t+n}, a_{t+n}) = 0$$

と定義する．エピソード終了時には，

$$Q(s_{t+n-1}, a_{t+n-1}) = r_{t+n}$$

である．このようにエピソードを終了することで，1時刻前の価値が確定する．リターンが起こるまで待たなくて良いことになる．モンテカルロ法ではエピソード終了時のリターンRを目標にした．TD法の目標は$r_{t+1} + \gamma Q(s_{t+1}, a_{t+1})$である．式1.4において$R$を$r_{t+1} + \gamma Q(s_{t+1}, a_{t+1})$でおきかえる．

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha(r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)) \tag{1.5}$$

各種アルゴリズム

モンテカルロ法の説明のさいに，行動選択の方法について，環境の知識利用と探索の双対問題を考慮したものをいくつかあげた．これらを用いながら環境を探索し方針の評価を行う．これらは決定論的にgreedy actionを選択するのではなく，時としてnongreedy actionを選択する．式1.5はsarsaアルゴリズムの更新式である．実際に選択された行動の価値を用いて評価を行っている．この式は行動方針と評価方針が一致している．この場合on-policyとよぶ．

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha(r_{t+1} + \gamma \max_a Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t))$$

この式はQ-learningアルゴリズムの更新式である．選択された行動とは無関係に選択可能な行動の価値の中で最大のものを用いて評価を行っている．下の式は行動方針と評価方針が分離されている．この場合off-policyと呼ぶ．

TD 法には行動価値関数と方針関数を明示的に分離した構造になっている actor-critic 法がある．方針関数は actor である．価値関数は critic である．critic は actor の行動選択を知り，結果が期待より良かったか悪かったを評価するので，学習は on-policy である．評価値は critic からの唯一の出力で，図 1.3にしめすように，acttor と critic の学習を促進する．critic は状態評価関数である．行動選択後の新しい状態を評価し，行動選択が期待より良かったのか悪かったのかを決定する．この評価を TD エラーと呼ぶ．

$$\sigma_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$$

Vは critic が見積もった現在の価値関数である．時刻 t の TD エラーσ_t は，状態 s_tで選択した行動 a_tの評価に用いる．もし TD エラーが正なら a_tを選択する傾向が強めることを示唆している．逆に負なら選択する傾向を弱めることを示唆する．方針を Gibbs softmax 法により生成すると

$$\pi_t(s,a) = Pr\{a_t = a | s_t = s\} = \frac{e^{p(s,a)}}{\sum_b e^p(s,a)}$$

$p(s,a)$ は時刻 t での actor の方針を修正するパラメーターである．状態 s で行動 a を選択する傾向を示している．上記に述べた強弱の扱いは，$p(s_t,a_t)$ を増加あるいは減少させる．具体的には，

$$p(s_t,a_t) \leftarrow p(s_t,a_t) + \beta\sigma_t$$

βは正のパラメータである．TD 法を用いた初期の強化学習システムは actor-critic である．以後，多くの注意が行動価値関数の見積もりと方針の決定を学ぶことにささげられてきた．[30, 31, 32] しかし，Actor-critic は次の 3 点から興味深い．

- 行動を選択するために必要最低限の計算を必要とする．たとえば，行動が連続値である事例を考えると，無限の選択可能な行動がある．方針があらかじめ
- 推計的方針を学ぶことができる．すなわち，それらは最適な確率を学ぶことができる．この能力はマルコフ性を満たさない事例で役立つ．
- 生物学的立場に適用しやすい．

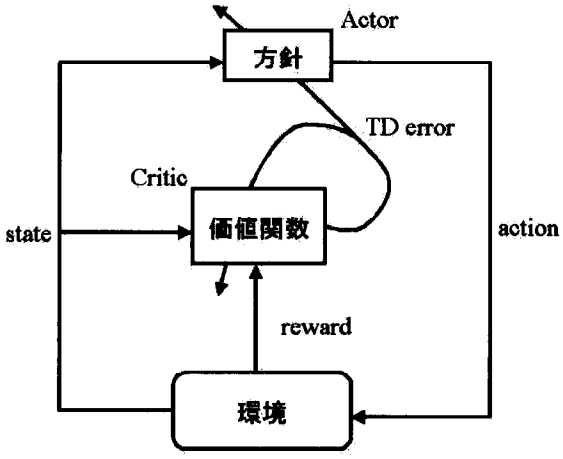

図 1.3: actor critic 構造

1.5 脳モデルとしての強化学習

基底核は，大脳と脳幹の間に位置し，大脳皮質や視床などの広範囲から入力をうけ，主な出力を視床経由で大脳前頭皮質に送りかえしている．大脳基底核の役目は大脳皮質機能のコントロールであるとされている．その中でも 2 つの大きな役割があると考えられている．そのひとつは運動や行動の選択に関わっていること，もうひとつは，運動計画の新たな獲得や行動の学習である．神経伝達物質であるドーパミンは大脳基底核でもっとも良く機能している．ドーパミンによるニューロン群の活動は神経回路による選択の役割を果たす．

シュワルツらは運動課題を遂行中のサルの神経活動記録によりドーパミン細胞が行動の結果得られる餌などの報酬にたいして，あるいは報酬を期待させるような信号にたいして応答することが明らかにされた [23]．シュワルツら [23] の動物実験がしめすドーパミン細胞の働きと actor critic モデルはよくにている．ドーパミン細胞の応答特性は TD エラーの応答そのものである．この事実に触発され Houk,Barto は [29]actor critic モデルが大脳基底核にインプリメントされているという仮説を発表した．これは脳モデルに対する強化学習の応用例の一部である．強化学習がたんに理論的可能性をあらわすだけでなく脳内にあるという予想にもとづき研究が続けられている．

第 2 章

モデル

2.1 能動的認識の図式化

能動的認識の枠組を図 2.1に示す．このモデルは対象を観測する感覚部と，これに基づき対象の内部像を形成する認識部から構成される．

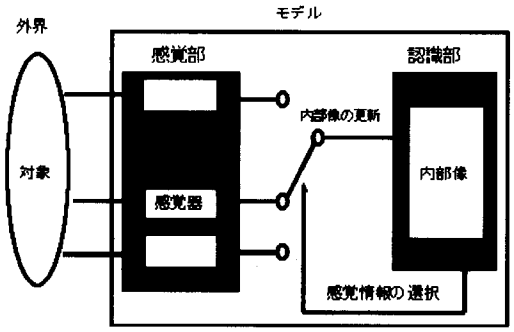

図 2.1: 単純化した能動的認識

認識部は感覚器を逐次的に選択して，それらを統合し対象の《内部像》を構成する，内部像が明確になるまで，感覚器の選択，対象の観測，内部像の更新が繰り返される．ある時点で感覚器を選択するかについては，その時点での認識の進み具合にに依存する．これにより能動的認識の状況依存性機能が実現される．このことは初期状態での先入観をを仮定することで文脈を反映した認識も実現可能になる．こうした選択機能を実現するためには,《ある対象をある感覚器で観測すればどのような信号が得られるか》という外界事象に関する知識をなんらかの形で認識部にたくわえておく必要がある．

- 対象に関する部分的な感覚情報が逐次的に統合されて内部像が構成される.
- 各時点で内部像に応じて感覚器が選択される.
- 適切な感覚器選択には外界の情報構造に関する知識が必要となる.

能動的認識の学習には2つの側面がある. ひとつは感覚器の選択方法に関する学習である. 能動的認識では感覚情報を逐次的に統合した結果として認識の結論が得られるので, 全ての情報を一括して処理する受動的な認識系とは異なる問題が生じる. ひとつは, 感覚器を選択した時点でその選択が良いのか悪いのかが明らかでないことである. 認識の結果は一連の感覚器利用後に与えられるので, 選択に対する評価は結論を出した時点まで得られないから, どの感覚器の利用を学習した良いかが明らかではない. この問題は, 成功への保証がどのように配分されるかという問題で古くから指摘されてきた [16]. この問題に対する解決方法として注目をあびているのが, 行動系列学習などに用いられることの多い強化学習の一種である TD 学習である [?]. この学習方法は, 試行錯誤により各時点での報酬の予測をおこない. 予測誤差を起因とした学習をおこなう. この仕組みを採用する. したがって, 報酬の予測と予測誤差を生成し疑似的に報酬信号を生成する《報酬系》をあらたに設ける.

もう一つの問題は探索空間の広さである. 利用できる感覚器の数が膨大な場合は試行錯誤だけで有用な感覚器を捜し出すことは実質的に不可能であり, 効率的な感覚器選択が実現できない. 本研究ではとくに対策を行わず試行錯誤にまかせることにしている.

感覚器の選択とは別にもうひとつ学習がある. 外界事象に関する知識の学習である. ある対象を, ある方法で観察すれば, どんな信号が得られるかについての学習である, 現在の観察信号がいったいどの対象に属するかを決める問題である. 解決方法のひとつは, 観察信号をどの対象に含めるかを試行錯誤することである. この際も感覚器選択と同様の問題が生じるが, 同様に解決することにする.

以上の考察から図 2.1のモデルを改変したものを図 2.2に示す. このモデルでは, 図 2.1のモデルの感覚部を感覚系, 認識部を認識系と呼び, 認識系の出力に応じて感覚器の選択を行う部分と認識結果を出力する部分を選択系として独立させた. 認識の目的はすでに与えられているとし, 認識結果にたいして目的達成を通知する教師を配置した. また, 学習

を行う上で必要な学習信号を生成する報酬系を新たに加えた．認識結果出力の後，学習者は報酬系において，教師からの通知をうけとり，内部で学習信号を生成する．目的達成を増加させるために学習信号を認識系，選択系に送る．

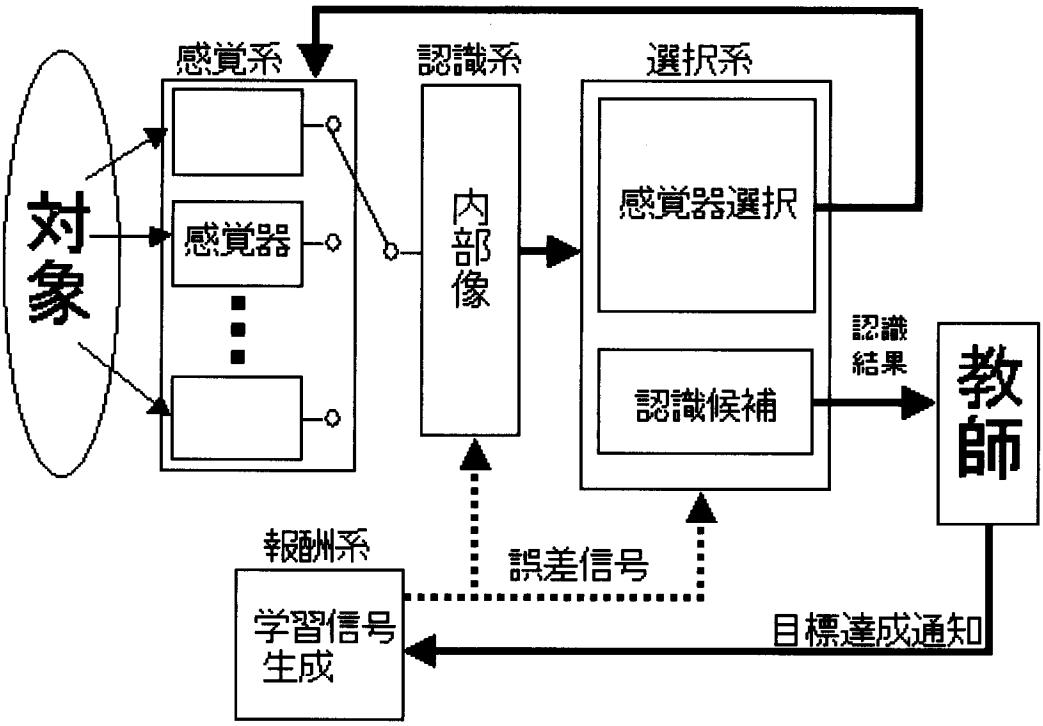

図 2.2: 学習を考慮した構造

2.2 構築したモデルの構造とアルゴリズム

実際に構築したモデルについて述べる．能動的認識のモデル内には外界の観察結果を内部像に変換する認識系がある．内部像に基づいて感覚器を選択する選択系がある．これらの機能の学習に寄与する報酬系がある．まず能動的認識を実現する機能である認識系と選択系について述べ，報酬系については，これらの学習とともに述べることにする．

2.2.1 モデルの構造と入出力関係

認識系は観察信号を入力とし内部像を出力とする関数である．選択系は内部像を入力を受け，選択（感覚器選択あるいは認識結果出力）を出力する関数である．認識系，選択系の関数機能を実現する方法に Radial Basis Function Netowork(RBFN) を用いた．RBFN は Basis Function(基底関数) により入力信号を特徴化し，出力を基底関数を用いて線形近似するものである．この仕組みを用いると，入力信号ベクトル x にたいする基底関数を

$$\phi_j^x = \exp\left(-\frac{||x-c_j||^2}{2\sigma^2}\right)$$

で定め，出力ベクトル $y=(y_1,y_2,...,y_N)$ を

$$y_k = \sum_{j=0}^{N} \lambda_{kj}\phi_j^x$$

で表すことができる．RBFN の概念図を図 2.3にしめす．$j(j=1,2,...,N)$ は中間ノード番号あらわしている．中間ノードの出力が基底関数の出力である．入力信号 x を N個の中間ノード $(\phi_j^x(j=1,2,..N))$ でうけ，出力ノード yへ送っている．中間ノードの応答特性を図 2.4にしめす．j番目の中間ノードは，そのノードパラメータ c_jからの距離に応じて，距離が近いほど 1 に近い値を出力する．すなわち，特定の入力信号に応答する素子である．そして出力ノードに接続するさいの荷重 (λ_{jk}) は，中間ノード jの影響を出力ノードへ伝える役目をする．σは曲線の傾きを決定するパラメータである．基底関数の数にあわせて定める必要がある．RBFN は学習信号を生成する報酬系でも用いる．

認識系

認識系には，認識対象ごとに認識素子を配置する．時刻 t に，認識素子 $n(n=1,2,\ldots,N)$ は選択された感覚器 $s(s=1,2,\ldots,S-1)$ により得た信号 x_sをうけとる．この信号によっ

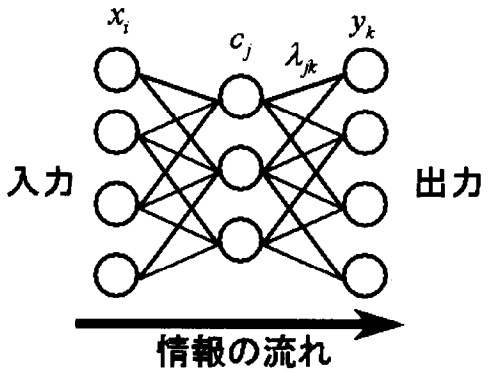

図 2.3: RBFN 概念図

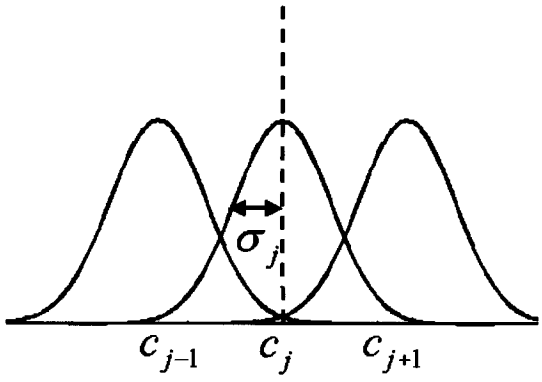

図 2.4: 基底関数応答特性

て活性化する量を $u_n(t)$ であらわす．この活性化量を蓄積したものを $p_n(t)$ であらわす．ベクトル p が内部像である．内部像に基づき認識候補を定める．どの対象を候補に選ぶかは各素子の活性化量 $p_n(t)$ の大きさに比例した確率で選ぶ．

$$\pi(p_n(t)) = \frac{p_n(t)}{\sum_{k=1}^{N} p_k(t)}$$

この認識候補は後に選択系で認識結果の出力が選択された場合に出力されることになる．認識素子を活性化する仕組みには RBFN を用いた．入力信号 x_s にたいする基底関数を

$$\phi_i^x = \exp\left(-\frac{||x - c_i^x||^2}{2\sigma^2}\right)$$

で定め，n 番目の認識素子の活性化量を

$$u_n(t) = \sum_{i=0}^{N} o_{ni}(t)\phi_i^x(t)$$

で表す．入力信号は時刻 t に選択した感覚器 kからのみである．$i(i = 1, 2, \ldots, I)$ は基底関数の番号である．こうして得た活性量は出力ノードにて

$$p_n(t) = p_n(t-1) + u_n(t)$$

により時系列的に保持される．

RBFN の役割は，認識素子が適切に興奮できるしくみを作ることである．中間ノードは観察信号を特徴化する．認識素子が，どの中間ノードからの信号を重視するかを過重 o_{ni}で決定する．ある感覚器を用いると，どんな信号が観察されるか？これを基底関数で表現し，そしてそれらがいずれの認識対象にふりわけられるのか？を過重で表現する．

選択系

選択系には，選択肢ごとに選択素子を配置する．時刻 t に，選択素子 $s(s = 1, 2, .., S)$ は認識系より出力された内部像ベクトル p をうけとる．この信号による素子の活性化量を $a_s(t)$ であらわす．選択の決定は各素子の活性化量 $a_s(t)$ の大きさに比例した確率で選ぶ．確率をπで表す．

$$\pi(a_s(t)) = \frac{a_s(t)}{\sum_{k=0}^{N} a_k(t)}$$

この選択肢には感覚器選択，認識系で定めた認識結果の出力が含まれる．$s = 1, 2, ..S-1$ にたいしては感覚器を割り当て，$s = S$には認識結果の出力を割り当てる．選択素子が活性化する仕組みにはRBFNを用いた．入力信号ベクトルpにたいする基底関数を

$$\phi_j^p = exp\left(-\frac{||p - c_j^p||^2}{2\sigma^2}\right)$$

で定め，s番目の選択素子の活性化量を

$$a_s(t) = \sum_{i=0}^{N} w_{sj}(t)\phi_j^p(t)$$

で表す．$j(j = 1, 2, .., J)$は基底関数の番号である．認識系では時刻tの入力信号は選択した感覚器kからであったのにたいして，選択系ではすべての認識素子から入力を受ける．

RBFNの役割は認識の進み具合に応じて選択を決めることである．中間ノードの役割は認識の進み具合を特徴化することである．内部像ベクトルpを構成する認識素子$p_n(t)$の活動は，拮抗しているときには低く，候補が絞られてくると高い活動をするものが現れてくる．認識素子の活動レベルに応じて，異なる中間ノードが強く応答する．選択素子$a_s(t)$は，これらの中間ノードϕ_p^jから信号をうけとり，活性化することになる．選択素子が，どの中間ノードからの信号を重視するかは過重w_{sj}で決定する．

2.2.2 学習のアルゴリズム

学習とは選択系，認識系で設定したパラメータを目標達成にたいして適正なものにすることである．そのために，モデルは目標達成の予測を行っている．この予測の誤差に基づきパラメータを修正する．予測をおこなっているのが報酬系である．報酬系の目標達成の予測と予測誤差生成，そしてこの誤差に基づいた認識系，選択系，報酬系自身のパラメータ修正について述べる．

報酬系

報酬系には，予測素子を配置する．時刻tに，予測素子は認識系より出力された内部像ベクトルpをうけとる．この信号による素子の活性化量を$e(t)$であらわす．この予測は内部像変化の前後で行われる．報酬系は認識結果にたいする正否を教師から受け取る．報

酬系は，認識結果が正しい場合は，$R(t)=1$ を，間違えたり，認識結果を出さない場合には $R(t)=0$ を教師から受け取る．報酬系は学習信号を過去の報酬予測 $e(t-1)$，現在の報酬予測 $e(t)$，教師からの通知 $R(t)$ をもとに次式から計算する．

$$r(t)=R(t)+\gamma e(t)-e(t-1)$$

ここではγは $0<\gamma<1$ を満たす定数である．$r(t)$ は予測誤差をあらわす．これを認識系，選択系に送る．報酬系自身のふるまいも，この予測誤差をもとに修正する．

次に予測について述べる．予測は予測素子の活性化量であらわす．予測素子が活性化する仕組みには RBFN を用いた．入力信号ベクトル p に対する基底関数は選択系の場合と同様に

$$\phi_i^p=exp\left(-\frac{||p-c_i||^2}{2\sigma^2}\right)$$

で定め予測素子の活性化量を

$$e(t)=\sum_{i=0}^{N}v_i(t)\phi_i^p(t)$$

で表す．

中間ノードの役割は認識の進み具合に応じて認識達成の予測を決めることである．内部像 p を構成する認識素子 $p_i(t)$ の活動は，拮抗しているときには低く，候補が絞られてくると高い活動をするものが現れてくる．認識素子の活動レベルに応じて，異なる中間ノードが応答する．予測素子 $e(t)$ は，これらの中間ノード ϕ_p^i から信号をうけとり状況に応じて活性化することになる．そのさいの活性量は，認識候補が拮抗しているときは低く，候補が絞られるにつれて高くなる必要がある．予測素子は各中間ノードからの信号を重視するかは過重 v_i で決定する．

予測誤差に基づいて認識系，選択系，報酬系の振る舞いを修正することについて述べる．2つの学習がある．ひとつは基底関数の修正である．もうひとつは基底関数から各素子への接続荷重である．基底関数は入力信号の特徴をとらえることである．予測誤差は基底関数が適切であるか否かをあらわす．基底関数が入力信号にたいして適切になるように修正する．具体的には c_j を変化させることで基底関数の位置を移動させる．観察信号 x_k にたいする j 番目の認識系の基底関数の修正は

$$c_j^x(t)=c_j^x(t-1)+(x_k-c_j^x)\phi_j^x(t)r(t)\eta$$

で行う．内部像ベクトル p にたいする j番目の選択系，報酬系の基底関数の修正は

$$c_j^p(t) = c_j^p(t-1) + (p - c_j^p)\phi_j^p(t)r(t)\eta$$

で行う．

次に基底関数から各素子（認識素子，選択素子，予測素子）への荷重について述べる．まず，認識系にとって，予測誤差は認識素子の興奮が適切であったか否かをあらわす．$r(t)$ が正の場合，活性化促進を意味する．負の場合は活性化抑制を意味する．零の場合は活性化の程度が適切であることを意味する．

$$o_{ji}(t) = o_{ji}(t-1) + r(t)\phi_j^x\eta$$

ここで i は時刻 $t-1$ の認識候補である．

選択系にとって，予測誤差は選択が適切であったか否かをあらわす．$r(t)$ が正の場合，選択促進を意味する．負の場合は選択抑制を意味する．零の場合は選択が適切であることを意味する．中間ノードから選択素子への荷重を

$$w_{ji}(t) = w_{ji}(t-1) + r(t)\phi_j^p\eta$$

により修正する．ここで i は時刻 $t-1$ の選択である．

報酬系にとっての予測誤差は予測が適切であったか否かは表す．$r(t)$ が正の場合，予測より結果が良かったことを意味する．負の場合は予測より悪かったことを意味する．零の場合は予測が適切であることを意味する．中間ノードから予測素子への荷重を

$$v_j(t) = v_j(t-1) + r(t)\phi_j^p\eta$$

により修正する．

2.3 実験：認識系の学習を完了済とした場合

2.3.1 目的

簡単なパターン認識を例題としてモデルの振舞いを検討する．モデルの要となる学習のうち認識系学習は完了済として検討する．

2.3.2 方法，条件

- 刺激

  ４×４の２値パターンを取り扱う問題を設定した．認識すべきパターンは図 2.5に示す６つを用意した．使用可能な感覚器は図 2.6に示す受容野をもつ２０個を用意した．各番号は感覚器の種別を表す．各感覚器は自分の担当する受容野内の濃度平均を出力する．

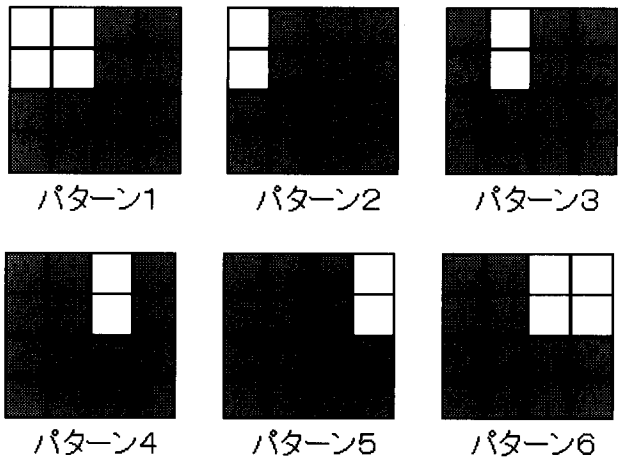

図 2.5: 例題のパターン

- 手順

  パターンが提示されると，モデルは認識結果を出力するまで観察を行う．これを一試行とする．一回の観察で，一つだけの感覚器を使用する．必要ならば複数回数の

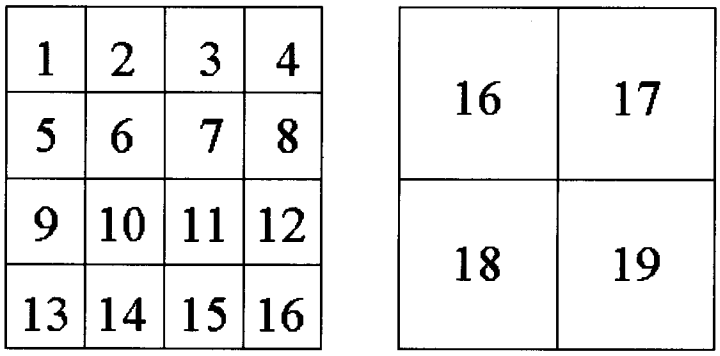

図 2.6: 感覚器の受容野

観察を行うことになる．パターンは各試行ごとにランダムに提示される．認識結果に対して正否が与えられる．認識結果が正しい場合には正の値を報酬として与える．間違っていた場合は負の値を罰として与える．認識結果を出力しない場合は零値を与える．

2.3.3 結果

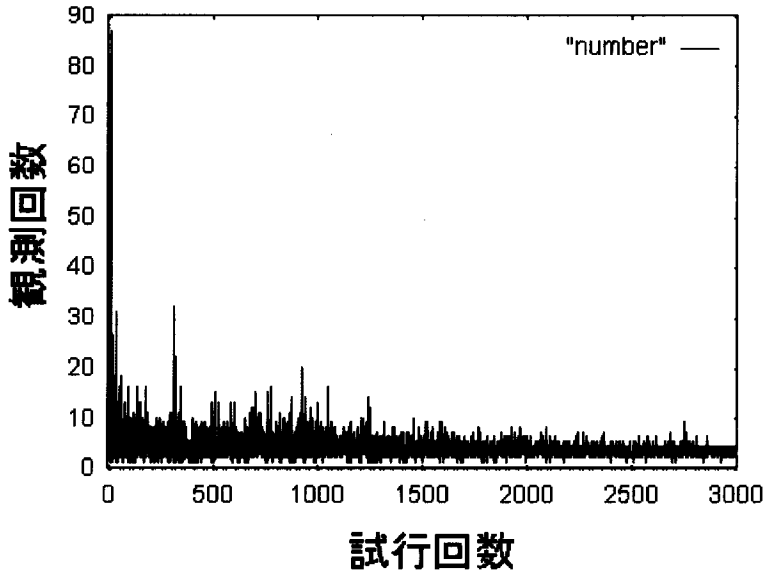

図 2.7: 実験結果:試行に伴う平均観察回数の変化

図 2.7に試行回数と観測回数の関係示す．一試行とはパターンが提示されて，モデルが必要とした観察を行い，認識結果を出力するまでを指している．

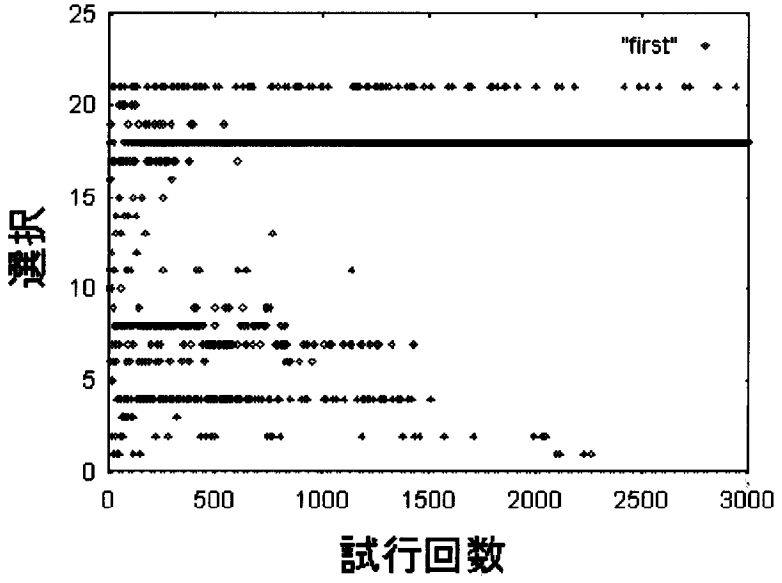

図 2.8: 実験結果:パターン提示後，第 1 回目の選択

図 2.8に示すのはパターンが提示されて一度目に行った選択の分布を示している．学習の初期は様々な選択を試している．しかし，1500 回を過ぎると 1 つの選択に絞られてくる．

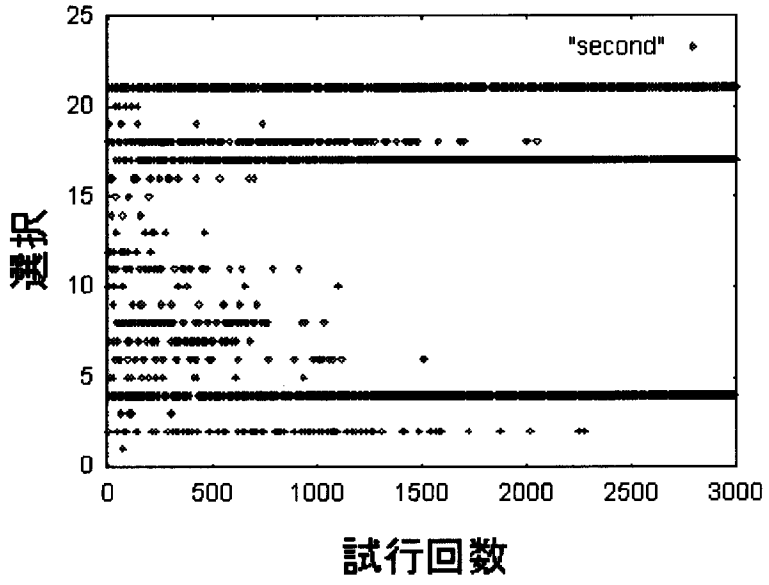

図 2.9: 実験結果:パターン提示後，第 2 回目の選択

図 2.9に示すのはパターンが提示されて二度目に行った選択の分布を示している．学習の初期は様々な選択を試している．しかし，一度目と異なるのは，終盤になって 2 つの選択に絞られていることである．これは一度目の選択の後，状況に応じた 2 つの選択が残ったと考えられる．

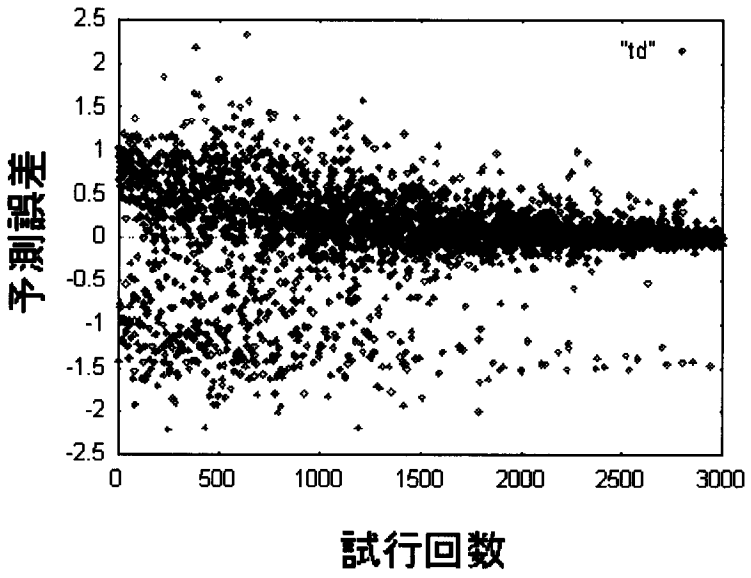

図 2.10: 実験結果:予測誤差の推移

図 2.10に示すのは予測誤差の推移である．試行回数が増えるにつれて予測誤差は零に集まっており学習の効果がみられる．モデルは認識過程の各状態で認識結果の正しさの観点から評価をおこなっており，これの誤差をあらわす．

2.3.4 考察

試行回数が増えるにつれ，認識結果を出力するまでの観測回数が減少した (図 2.7)．感覚器 17 あるいは 18 を用いた大局的な観察の後，内部像に応じた感覚器 1～8 の選択が見られた．

この結果の一つの解釈として，実験に用いた課題は図 2.11に示す木構造を持っているので，この木構造をたどることで効率的に認識が行われる．と考えられる．図 2.11の丸の中に数字が記してあるのは選択した感覚器を表す．丸がついていないものは認識結果である．獲得された感覚器の利用規則は課題の構造を反映するものと考えられる．利用規則の獲得は木構造の作成と見なせる．この問題では図 2.11に示すように複数の木構造が最適となる．このように問題に対して一意に木構造が決まらない場合が起こるが，これは問題自体の構造であり，アルゴリズムとは直接関係がない．

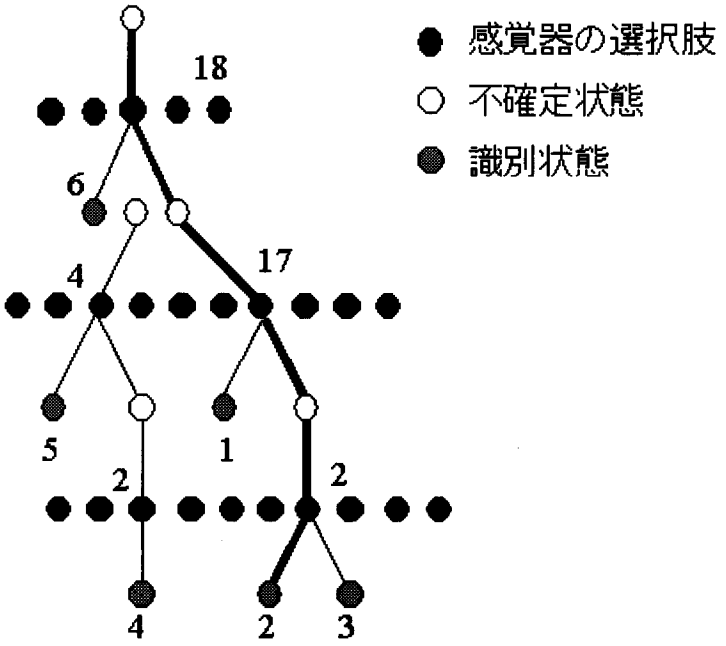

図 2.11: 課題のもつ木構造

第 3 章

一般的考察

3.1 学習の動機づけ：報酬信号の与え方

2つの学習を同時に行おうとすることに困難が生じた．その理由の一つとして，選択系の学習は，学習済の認識系学習の上になりたつためであると思われる．学習を誘導する報酬は源は同じで，ともにモデルの外部にある．しかし，逆のことも言える．すなわち，選択系の獲得する情報をもとに認識系学習が行われるため，認識系の学習は選択系に基づいていると見なせることもできる．両者の学習の動気づけを分離して競合関係に置くことが良いと思われる．今後，2つの学習を同時に行う仕組みを考案する過程で，この問題の解決に関しては，適切な実験方法を考慮した上で，明らかにしていく必要があろう．

本研究では学習を動機付ける要因として，外部から認識結果の正否を明示的に与えた．結果が良かったのか悪かったのかだけを与えることは強化学習では一般的な方法である．ここで外部と呼んでいるのは，エージェントの外部であって学習者の外部ではない．学習者は内部に動機付けをする教師をもち，学習者内の認識エージェントを制御する．それでは，学習者は学習前に答えを知っているのだろうか？

認識の目的は，個体の目的となんらかのむすびつきがある．ある状況で何をすればいいのか知りたいならば，今，どういう状況かを区別しなければならない．たとえば，病気をしているときに薬を飲みたいならば，目の前にある瓶のラベルに表記されている毒あるいは薬という文字を区別できなければ，目の前のどちらの物を口にしたらよいのかわからない．認識結果を得るために状況に応じた感覚器選択を結びつけたように，正しい認識結果を得ることが後の大きな目的にむすびつく．したがって，この場合，認識結果は最終目標ではなく，ひとつのサブゴールにすぎない．しかし，本研究ではサブゴールとして認識の

目標をを作り出す部分は扱わずに，認識の目標が定まった状態から出発して取り扱っている．人間には，どんな動機づけがあるのか？もっともわかりやすいのは本能である．しかし，人間は社会的存在でもあるから，社会性に依存した動機付けもあると思われる．

もうひとつの形態として学習エージェントが自発的に《あるものと，また，別のあるものを区別する》というふうに，区別を始めることも考えられる．普段の生活でも，この場合，認識エージェントには報酬が外部から明示的に与えられない．つまり，明確な目標にむすびついていない場合である．明確な目的と結びついて動機付けられる場合と，その結びつきのない場合があり，後者の場合は，認識エージェントが固有にもっている弱い動機づけもあるように思われる．

人間の活動には複数のエージェントが存在しており認識エージェントもその中の一つである．複数のエージェントは独立に存在していることもできるが，特定のエージェントが強く活動する場合には，周囲のエージェントは協力するようになっていると思われる．たとえば生体の危険を知らせるエージェントが強く活動すると，危険を回避するように認識エージェントは，協力を要請される．認識エージェント固有に存在した弱い動機付けは，危険を回避するという強い動機付けにより相対的に弱められてしまう．生体の危険を回避するエージェントの活動が弱まるまで協力は続く．このようにエージェントの学習の動機付けは，生体の目的の明確さに関わってくると考えられる．今後，認識エージェント固有の動気づけにともなう学習と認識エージェント以外の目的に結びつけられる学習の仕組みに発展することが求められる．

3.2 逐次的学習

認識学習のモデルは，通常，認識すべき対象の集合が予め設定されており，その対象の中から繰り返し提示することにより，次第に認識能力が向上するという定式化を行うことが多い．しかし，人間が新しい対象の認識能力を次第に拡張していくことを考えると，この定式化は誤った形でとらえていることになる．人間の認識学習のモデルを構築するには，新しい問題をひとつずつ学習する逐次的学習を取り入れる必要がある．最初は単純な問題から始め，次第に複雑な問題の学習へ進んでいく状況を設定するべきであると考えられる．

能動的認識の学習は，感覚器選択の木構造を作り出す過程とみなせる．多数のパターンを同時に学習することは，複雑な木構造を一度に作ることに相当する．それより，簡単な土台となる木構造から作り，次第に枝を増やす手順を踏めば，複雑な木構造へと成長させることに無理はない．つまり，その時点で認識能力適した問題を順次学習させるのが効率適である．あらかじめ利用可能な感覚器を限定しているが，感覚器利用のレパートリーを増加させることも望まれる．

また，本研究のモデルでは識別対象ごとに，あらかじめ認識素子をわりあてている．しかし，まだ存在の知らない対象に素子が割り当てられており，認識能力が低いとき，すなわち識別対象が少ないときと多いときでは素子の数も異なっているほうが自然である．あらかじめ識別可能な対象に限りがあるのは能力を拡張する逐次的学習を考慮すれば，認識素子をモデル自身が必要に応じて増やす仕組みも望まれる．

3.2.1 教師の役割

単純な課題から複雑な課題へという逐次的学習を行うためには，学習課題の統制を行う必要がある．課題の学習順序は認識系を成長させていくさいに重要な意味をもつ．特殊な課題を学習させたければ，その問題に特異的な感覚器選択を取らせることが効率的である．しかし，そのさいのクセは新たな学習課題を達成するさいに障害となる．つまり，いったん特殊な木構造を作ると別の木構造を作り出すことが困難になるわけである．

モデルの検証に用いた例題では，複数の木構造のうちどれを作り出しても問題はなかったが，複雑な課題にたいして逐次的学習により木構造を成長させるためには，最初にどのような木構造を作るかが重要であると考えられる．

こうした統制を行う存在として教師があげられる．教師は単に認識結果が正しいか否かのみならず，次に何を学習したらよいかという課題の設定を行う存在として重要であると考えられる．

3.2.2 モデル自身による問題の取捨選択

問題の提示順序の統制を教師にまかせるのも一つの方法である．しかし，モデル自らが行うメカニズムを埋め込むことも興味深い．モデル自身が認識能力の発達段階に応じて問

題を選ぶしくみである．ここで問題を選ぶとは，難しいと感じた問題はすぐにあきらめることに相当する．この機能実現は次の二点から考えることができる．

ひとつは利用する感覚器の種類が増えるにつれて獲得されるというものである．特異的な効果をもつ感覚器は，認識能力の低い段階では使いこなすことができずに，利用をあきらめる．この段階では，どの対象にも汎用性のある感覚器を利用している．このような感覚器の利用では，対象をあらわす木構造の根付近しか構成できない．この根付近は，簡単な課題に相当する．

もうひとつは，個別の問題を解く目的だけでなく，複数の課題を解くさいの全体的な目的を与えることである．全体目的とは一定期間の報酬の蓄積ができるかぎり多くすることである．一定期間の報酬をあらわすものの一つは，正解数／総問題数 である．正解率である．もうひとつは, 正解数／総手続き数 である．これは，認識の効率を表す．2 値の合計値を指標に問題への見切りを決める．この値は正当率と認識効率のバランスから成り立っている．どちらを重視するかは状況による．手続きの数が同じ問題を同レベル問題とすれば，あるレベルをマスターするまでは，そのレベルに必要な手続き回数内に絞り，さらに高いレベルに取り組む場合は問題に費す時間を増やすことなる．このようにして一定期間の報酬蓄積増大という目的を遂げるために問題に費す時間を調節すると思われる．

このように自律的な逐次的学習は，必要な情報を選択 (選択には無視も含む) する能動的認識の特徴のうえに成り立つと考え，本研究のモデルはこれらを実現する基盤となるものと考えたい．

3.3 今後の展開

本研究は，能動的認識に着目して認識のモデルを構成した．2 つの実験から学習を検討した．モデルのメカニズムについて考察を行ってきた．前述してきたような目標達成に動機づけられた学習の傾向が得られた．しかし，2 つの学習をゼロからスタートとすると，学習進行に困難が生じた．選択系の学習は認識系の学習の基づいており，学習の報酬源を別にしたモデルを構築することを考える．その後に，能力拡張の逐次的学習やモデル自身による問題の取捨選択に取り組む．

第 4 章

結論

本研究では認識のモデルを能動性に着目して定式化した．本モデルは大きく 2 つに分けられる．一つは認識系である．これは従来より扱われてきた認識の機能にあたる．受動的に与えられた情報を処理する．もうひとつは選択系である．これは処理する情報を選択する機能である．

人間は能動的認識を誰からか逐一教わったというよりは自己組織化したと思われる．自己組織化の仕組みとしては，認識に役立った情報は再び選択されやすくなったと考え，学習の枠組みに強化学習を用いた．本モデルは認識結果の正否だけが与えられる学習状況を想定する．能動的認識には情報が逐次的に選択されるゆえ起こる課題がある．認識結果を出力する直前の情報選択については，認識結果の正否から明示的に強化が可能である．しかし，それ以前の情報選択については明示的でないため，みずから試行錯誤の中，評価しなければならない．この問題の解決法として強化学習の一種である TD 学習を用いた．モデルを，簡単な例題を設定し，計算機実験により検討した．選択系の学習につていは動作を確認した．

今後，認識系学習について検討するとともに，人間の振舞いとモデルの振舞いを比較するための課題の選定が必要である．モデルが，こうした課題を処理するさいに鍵となるのは，状況に応じて必要な情報を順次処理していくという能動的認識の特徴がより大きな課題を解くさいの手順の構築として寄与すると思われる．

謝辞

本論文の作成にあたり，御指導頂きました主任指導教官である阪口 豊 助教授に感謝いたします．ヒューマンインターフェース学講座の皆様に感謝いたいします．

参考文献

[1] 波多野完治 (編) : ピアジェの発達心理学, 国土社,1965.

[2] 竹田青嗣 : 現象学から実存主義へ, 現代思想入門,JICC 出版局,1984

[3] 石川正俊, 山崎弘郎 (編) : センサフュージョン, コロナ社,1992

[4] Gibson, J. J.: The ecological approach to visual perception, Houghton Mifflin, 1979.

[5] Neisser, U.: Cognition and reality, Freeman, 1976.

[6] Thorndike, E. L. (1911). Animal Intelligence. Hafner, Darien, Conn.

[7] Bellman, R. E. (1957a). Dynamic Programming. Princeton University Press, Princeton, NJ.

[8] Bellman, R. E. (1957b). A Markov decision process. Journal of Mathematical Mech., 6:679–684.

[9] Turing, A. M.(1950). Computing machinery and intelligence, Mind, 59:433-460. Reprinted in E.A. Feigenbaum and J.Feldman(eds),Computers and Thought, pp. 11-35. McGraw- Hill NewYork, 1963

[10] Minsky, M. L. (1954). Theory of Neural-Analog Reinforcement Systems and its Application to the Brain-Model Problem. PhD thesis, Princeton University

[11] Farley, B. G. and Clark, W. A. (1954). Simulation of self-organizing systems by digital computer. IRE Transactions on Information Theory, 4:76–84.

[12] Clark, W. A. and Farley, B. G. (1955). Generalization of pattern recognition in a self-organizing system. In Proceedings of the 1955 Western Joint Computer Conference, pages 86–91.

[13] Michie, D. (1963). Experiments on the mechanisation of game learning. 1. characterization of the model and its parameters.Computer Journal, 1:232–263.

[14] Michie, D. (1974). On Machine Intelligence. Edinburgh University Press.

[15] Widrow, B., Gupta, N. K., and Maitra, S. (1973). Punish/reward: Learning with a critic in adaptive threshold systems. IEEE Transactions on Systems, Man, and Cybernetics, 5:455–465.

[16] Minsky, M. L. (1961). Steps toward artificial intelligence. Proceedings of the Institute of Radio Engineers, 49:8–30. Reprinted in E. A. Feigenbaum and J. Feldman, editors, Computers and Thought. McGraw-Hill, New York, 406–450, 1963.

[17] Shannon, C. E. (1950a). A chess-playing machine. Scientific American, 182:48–51.

[18] Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal on Research and Development, pages 210–229. Reprinted in E. A. Feigenbaum and J. Feldman, editors, Computers and Thought, McGraw-Hill, New York, 1963.

[19] Sutton, R. S. (1988). Learning to predict by the method of temporal differences. Machine Learning, 3:9–44.

[20] Sutton, R. S. and Barto, A. G. (1981a). An adaptive network that constructs and uses an internal model of its world. Cognition and Brain Theory, 3:217–246.

[21] Sutton, R. S. and Barto, A. G. (1981b). Toward a modern theory of adaptive networks: Expectation and prediction. Psychological Review, 88:135–170

[22] Tesauro, G. J. (1992). Practical issues in temporal difference learning. Machine Learning, 8:257–277.

[23] Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. Science, 275:1593–1598.

[24] Watkins, C. J. C. H. (1989). Learning from Delayed Rewards. PhD thesis, Cambridge University, Cambridge, England.

[25] Witten, I. H. (1977). An adaptive optimal controller for discrete-time Markov environments. Information and Control, 34:286–295.

[26] Klopf, A. H. (1972). Brain function and adaptive systems—A heterostatic theory. Technical Report AFCRL-72-0164, Air Force Cambridge Research Laboratories, Bedford, MA. A summary appears in Proceedings of the International Conference on Systems, Man, and Cybernetics, 1974, IEEE Systems, Man, and Cybernetics Society, Dallas, TX.

[27] Singh, S. P. and Bertsekas, D. (1997). Reinforcement learning for dynamic channel allocation in cellular telephone systems. In Advances in Neural Information Processing Systems: Proceedings of the 1996 Conference, Cambridge, MA. MIT Press.

[28] Crites, R. H. and Barto, A. G. (1996). Improving elevator performance using reinforcement learning. In D. S. Touretzky, M. C. Mozer, M. E. H., editor, Advances in Neural Information Processing Systems: Proceedings of the 1995 Conference, pages 1017–1023, Cambridge, MA. MIT Press.

[29] Barto, A. G. (1995a). Adaptive critics and the basal ganglia. In Houk, J. C., Davis, J. L., and Beiser, D. G., editors, Models of Information Processing in the Basal Ganglia, pages 215–232. MIT Press, Cambridge, MA.

[30] Ant-Q: A Reinforcement Learning Approach to the Traveling Salesman Problem , by L.M. Gambardella and M. Dorigo,Universite Libre de Bruxelles, Bruxelles, Belgium.

[31] Jing Peng and RonaldJ. Williams. Incremental multi-step Q-learning. In Proceedings of the Eleventh International Conference on Machine Learning, pages 226-232, San

Francisco, CA, 1994. Morgan Kaufmann.

[32] R. Munos and J. Patinel : Reinforcement Learning with Dynamic Covering of State Action Space : Partitioning Q-learning, Animats, 3,355/363 (1994)

研究発表

[1] 高野光雄, 阪口 豊 “強化学習を用いた能動的認識の自己組織化モデル,” 電子情報通信学会ニュラールコンピューティング研究会, NC97-121(1998-03)

[2] 高野光雄, 阪口 豊, 出澤正徳 “強化学習の基づく能動的認識の自己組織化モデル,” 電気通信大学大学院, 第 6 回 IS シンポジウム, Sensing and Perception, 予定 (1999)