第１章　序論

１.１ 自律学習システムとニューラルネット

近年、コンピュータを始めとする知能システムの進歩はめざましいものがあり、我々の生活の中でなくてはならないものとなっている。にもかかわらず、我々は、これらを使っていく上で、言われたことしかしない、融通が効かない等の印象を受ける。つまり、我々人間を始めとする生物と比較して、数値演算能力ははるかに勝るものの、柔軟性・適応性という面で大きく劣る。本研究では、生物のような柔軟かつ適応的な知能システムの構築を大きな目標として置いている。

我々生物と従来の知能システムは、いずれも何らかの入力を受けてそれに対して何らかの出力をするという構成になっている。しかし、入力から出力への処理方法の獲得が自律的な学習によるかどうかという点と連続値を主体とした処理かどうかという点で相違点があり、それが前述のような差異を生み出していると考えられる。以下、この２つの相違点から、自律学習の必要性とニューラルネット適用の有効性について述べる。

1.1.1 自律学習システム

処理方法の獲得という点に着目すると、従来の知能システムは、主に、図１.１に示したように、人間が入力から出力への処理方法を与えるという形で知能化されている知識付与型の知能システムである。一方、我々生物は、入出力を見ながら自ら学習していくことができる。これをここでは**自律学習**と呼び、このような機能を持つシステムを**自律学習システム**と呼ぶ。中でも特に、図１.２のように運動出力と感覚入力を持ち、外界とのフィードバックループを形成することにより、処理方法を獲得・更新すること（フィードバック学習[Okabe 88]）が重要な役割を果たす。

知識付与型知能システムでは、言われたことを忠実にこなすことは可能であるが、自ら学習する手段を持っていない。しかし、知能システムに対する要求は年々高度になる一方である。この要求を満足するような知識を付与するためには、非常に緻密かつばく大な量のプログラミングが必要になると考えられる。要求される機能のレベルおよびプログラミングの難しさを定量的に表すことは困難であるが、直感的には、例えば機能のレベルと与える知識量の関係がピラミッドの容積と高さのように、要求される機能のレベルが高くなればなるほど、レベルを上げるためのプログラミングはそれ以上に難しくなってくると考えることができる。特に適応ということを考えると、あらゆる場面を予め想定し、それに対しプログラミングしなければならず、大変困難である。従って、初期の頃の知能システムの進歩と比較して、今後の知能システムの進歩の速度は徐々に鈍ってくることが予測される。

一方、自律学習システムである我々生物は、前述のようにフィードバックループを有し、自ら行

動し、外界の状態をセンサによって取り込み、自らの動作が引き起こす外界の変化を知ることによって学習し、知識を獲得していくことができる。従って、機能は学習によって獲得できるため、予め与えなくてはならない情報は少なくても良く、その分、適応的であると言える。また、第三者がいなくても自ら行動できるため、それだけ多くの経験を積むことができるし、行動することにより学習を行うため、必要な学習を能動的に行うことができる可能性も有する。これらのことから、将来の知能システムの発展を考えると、今こそ自律学習システムに着目し、我々生物をお手本としてその能力を向上させていくことが必要と考え、本研究を進めて来た。しかし、どこまでが自律学習かと言っても、その境界は必ずしもはっきりしない。例えば、システムの処理に一つのパラメータを導入し、入出力を見ながら学習によってそのパラメータを変化させていくといった方法も広い意味で自律学習と言える。このような意味で、システムの処理における既定部分と可変部分の比が一つの自律学習の尺度となる。そこで、ここでは、できる限り設計部分を減らし、汎用的な学習によって多くの知識を獲得することを目指す。ただし、可変部分と言っても初期値を与える必要があり、そこに予め知識を埋め込むことは可能である。

知識付与と自律学習の中間的な位置づけのものとして教師あり学習がある。中でも、ニューラルネットの教師あり学習アルゴリズムであるバックプロパゲーション法（以下、ＢＰ法と省略）[Rummelhart 86]は有名である。これは、図１.３のように、人間が入出力サンプルのセットを与えることによって、入出力関係を学習させるため、直接プログラミングしなくても良いという利点があり、人間がプログラミングする際に不得意としたパターン認識などパターン情報の処理をサンプルから

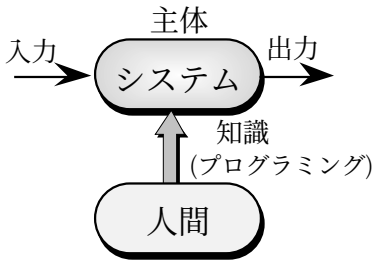

図１.１　従来の知能システム
（知識付与型知能システム）

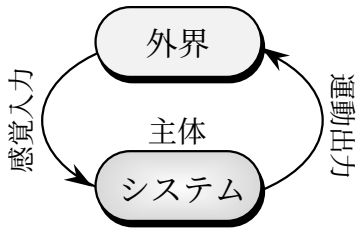

図１.２　生物を始めとする自律学習システム

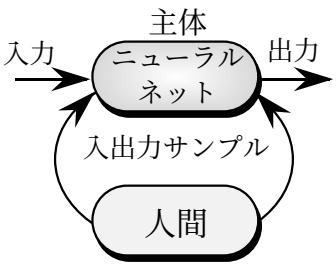

図１.３　バックプロパゲーション法によるニューラルネットの教師あり学習

の学習によって獲得できるということで威力を発揮した。そして、知識付与型知能システムからの脱却を夢見て多くの研究者が研究を行った。しかし、確かに直接知識を与えなくても良かったが、知識を入出力サンプルという形で付与しなければならなかった。しかも、そのサンプルを作ることが意外と難しく、特に、環境が変化していくような場合、そこにシステムが適応していくことも困難であるということがわかってきた。そして、何より、生物のように、フィードバックループを持っていないため、与えたサンプル以上の機能を学習によって獲得することはできないし、その機能が必ずしもそのシステムにとって最適なものであるかどうかわからないという、知識付与型知能システムと類似した壁にぶつかることになった。ここ数年、ニューラルネットの研究が少し下火になってきているが、この壁が下火になった大きな要因の一つと筆者は考える。

自律学習においては、いわゆるＢＰ法等の教師あり学習と違い、外部から直接理想出力である教師信号を得ることができず、自らの持つ何らかの学習指針に従って学習を進めていかなければならない。この学習の規準をどのようなものにするかが自律学習システムの性能を決定する大きなポイントとなる。自律学習は、外界とのループが重要な役割を果たすため、従来のように、文字認識、画像認識、音声認識、自然言語処理、制御といった個々の機能を独立して研究し、それを後で融合するというアプローチだけではなく、簡単でもいいから外界とのループを形成し、ループを利用した学習、ループを考慮した学習がどうあるべきかを考え、それを発展させていくというアプローチをとることが重要である。また、その学習は、構成の容易さという面からシンプルなほど良いが、学習の汎用性という面から考えても、シンプルな学習則からたくさんの機能を学習できる方がより柔軟性が高い学習則であると考えることができる。与える情報が多ければ多いほどそれに縛られ、学習による自由度が小さくなるからである。この関係を模式的に表したものを図１.４に示す。また、このような意味で、教師あり学習と比較すると、一般的に与えられる情報量が少ないため、より柔軟な学習が期待できる反面、探索空間が広くなるため、学習のために非常に長い時間を要するという欠点がある。

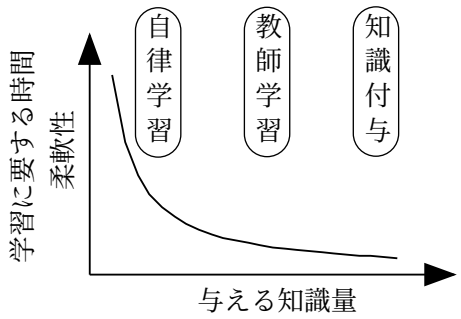

図１.４　与える情報量と柔軟性および学習に必要な時間の関係の模式図

1.1.2 パターン情報の学習・処理とニューラルネット

我々生物と従来の知能システムとの差異を内部での処理形態という点から捉えると、パターンを基盤としているか、シンボルを基盤としているかという違いが浮かび上がる。ただし、ここでパターンとは、距離および微分が意味をなす情報のこととする。

人間のような存在を考えると、一見、論理的思考、シンボル処理こそが高度な知能を支えているように見える。しかし、シンボルはその背後に隠れる様々な知識の象徴として存在し、その土台があるが故に、コミュニケーションや論理的思考といったシンボルの処理が有効になるものと考える。そして、その背後の知識がパターンとして構成されているため、距離が意味を持つことによって汎化能力につながり、微分情報を使ってその先を予測したり、山登り法（最急降下法）が適用できるため学習が効率的に行えるようになる。そして、これが我々生物が柔軟な処理を行うことができる大きな理由であると考える。シンボルとパターンの関係は、ちょうど図１.５のようにシンボルはちょうど氷山の一角のようなものであり、シンボルだけを取り出してその論理的な処理方法を考え、それを高次の処理と呼ぶことは、水面下の氷山の存在を無視して氷山のことを語ることに等しいと考える。

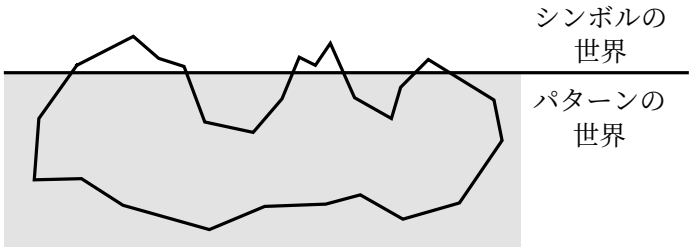

図１.５ 氷山を用いたシンボルとパターンの関係の模式図

パターン情報の処理を人間がプログラミングすることは、自由度が大きく、人間にとって大変困難な仕事となる。例えば、文字認識のプログラムを作ろうとしても簡単にはいかない。さらに、我々生物の処理を観察すると、様々な要因が複雑に絡み合って柔軟な処理をしているように見える。例えば、急いでキーボードに入力しようした時のミスの仕方を観察すると、キーボード上で近いキーを打ったり、発音が近いものを打ったり、単語の前後がひっくり返ったり、前後の文字の子音だけ入れ替わったり、近い意味の違う単語と間違えたりと様々な間違い方をすることにしばしば驚く。また、このような機能を人間がプログラミングしようとしてもほとんど不可能であると考えられる。従って、このような処理を実現しようとするならば､時間は掛かっても何らかの評価関数を作って微分情報を用いた学習に頼らざるを得ない。

以上の理由から、パターン情報の処理を得意とし、かつ学習能力を有するニューラルネットは自律学習システムを構築する上で強力な実現手段と考える。パターン情報を扱うニューラルネットの汎化能力は、過去の経験をいかに将来に生かし、柔軟な処理を行うことつながってくる。さらに、人間や他の生物といった自然界の高度な自律学習システムの中枢がニューラルネットで構成されており、言語などのシンボル処理ですらニューラルネットが担っていること、そして、この構成がおそらく長

い年月を経て最適化された結果であることもニューラルネットの使用の妥当性に関する大きな拠り所である。

また、ニューラルネットを使えば、複数の要因を適切にミックスすることが比較的容易にできる。例えば、筆者は以前、ロボットの制御を学習する際に、制御における誤差を小さくするという学習を行うだけで、フィードバック制御とフィードフォワード制御のハイブリッド制御の学習が可能であることを示した[柴田 89]。そしてさらに、ノイズが多い環境で学習した場合には、フィードバック制御が主流になるという適応能力も備えていることもわかった。このように、ニューラルネットは、学習のための評価関数を設定すれば、そのための適切な統合を学習によって獲得していくことができるのである。

一時期、右脳と左脳がパターン的な処理とシンボル的な処理を主に分担しているという知見を基に、それと同様にニューラルネットと従来の計算機を統合させれば人間のような柔軟なシステムができるといった議論がなされた。しかし、筆者は、全く異質な両者を最終的に統合することこそ難しく、統合することで両者の利点が失われる可能性が高いと考える。また、前述の議論のように、シンボル処理とパターン処理は非常に密接な関係にあるため、両者を分離するのではなく、一つの枠組みの中で捉えて行くべきであると考える。つまり、ニューラルネットを用いてシンボル処理を行うことによって、シンボルとパターンのインターフェイスの問題が解決され、ニューラルネットの学習・処理の柔軟さがシンボルを用いた表現の中にも十分に発揮されると考える。また、シンボル処理は、コミュニケーションや論理的思考を行うための必然から獲得できるという期待を持っている。

バックプロパゲーション法（ＢＰ）法は前節で述べたように、ニューラルネットを用いた教師あり学習の代表的なアルゴリズムであるが、ＢＰ法を単独で使うだけでは、教師信号を用意しなければならないため、自律学習を実現することはできないと述べた。しかし、 ＢＰ法は、教師信号と実際の出力の差の自乗を評価関数として最急降下法を適用した単純明快な学習アルゴリズムであり、中間層のニューロンを十分用意し、いくつかの教師信号を与えれば、元となった関数を近似することができるようになる等、その学習能力は強力なものであると筆者は認識する。そこで、本論文のほとんどの部分で、システムが教師信号を自動生成するという形で自律学習を保ちつつＢＰ法を適用するという手法をとった。ＢＰ法の具体的なアルゴリズムについては２.１節で、さらにＢＰ法を用いて強化学習を効率的に行う方法については、２.２節にて述べる。

ＢＰ法の前身である、パーセプトロンについては、小脳パーセプトロン説でも見られるように、それを実現するような機構が脳内に存在する可能性があると考えられている。しかし、ＢＰ法に関しては、逆伝搬誤差がニューロンをまたいで伝搬しなければならない等の理由から、そのような機構は脳内に存在しないと言う見方が大半である。しかし、私は、ニューロンの成長と学習を同一の範疇で捉えることができるのではないかと考えており、その考えに基づくと、神経の成長を促進する神経成長因子（NGF Neuro Growth Factor）[モンタルチニ 79][畠中 92] 等が学習にも関与し、さらには、ニューロンを介して他のニューロンへと伝搬することも考えられるのではないかと期待している。

また、ＢＰ法では、ニューロンの出力関数に非線形関数を用い、さらに、最急降下法を適用しているため、局所最適解（ローカルミニマ）に陥るという問題点がよく指摘されている。しかし、本論文の研究を進めるに当たって、ローカルミニマにトラップされて学習がうまく行かなかったという状況にはほとんど遭遇しなかった。ローカルミニマの例として、Exculsive-OR（排他的論理和）の学習

が挙げられる。この問題は、２次元の入力空間上の４点で教師信号が与えられ、入力空間を教師信号の値によって分離する際に線形分離ができない問題である。しかし、この問題でも中間層のニューロン数を増やせばローカルミニマを回避することができるし、現実に我々が存在する空間は、我々が意識している以上に線形性が強く、また、たくさんのデータを取得することによってこれらの問題をあまり考慮する必要はないと考える。

さらに、最近、ニューロンの出力関数に単調増加型のシグモイド関数ではなく、ガウス関数のようなＲＢＦ（Radial Basis Function）を用いた場合をよく見かける。これは、局所的な情報の和として出力を表現しようとしたもので、わかりやすく、また、ある入力における学習が他の部分に影響を与えない等という点から良く使われている。しかし、前述のように、我々の住む世界は線形性が高いものであり、シグモイド関数のような線形性の強い関数を使う方が効率的に状態を表現できると共に、汎化能力も活かせると考えられる。また、シグモイド関数は、入力が０に近いほど入力に対する出力の感度が大きくなる。これは、100.0 と 100.1 の違いよりも 0.1 と 0.2 の違いを大きく感じる我々の感覚と合っており、-∞から∞の入力を０から１の出力に変換する効率的な方法であると考える。以上より、本論文中では、従来通りのシグモイド関数を用いた単純なＢＰ法を学習に用いた。

また、よく、ニューラルネットは中がブラックボックスだから、たとえ学習によって機能が獲得できたとしても意味がないという議論がある。しかし、筆者は、人間や生物自体の処理は前述のように非常に複雑で表現することが困難なものであると考える。従って、生物の機能の本質を探るならば、また、生物のような柔軟な機能を実現しようとするならば、個々の機能を正面から考え、解析していくよりも、その学習方法を解析し、その実現を試みる方がより適切であり、近道であると考える。

１.２　自律学習における学習指針と強化学習

１.１節において、自律学習においては、いかなる学習指針を用いるかが１つのポイントであることを述べた。この学習指針は、汎用的であり、かつ、これによってシステム（エージェント）が有効に働くものでなければならない。以下では、筆者が最も有効と考える強化学習とその他の学習指針について述べる。

1.2.1 強化学習

強化学習とは、図１.６のような仕組みの下で報酬や罰といった強化信号から、より報酬を得られて、罰を避けられる動作を学習することを言う。基本的には、何らかの動作を行って、報酬が得られればその動作を強化し、罰が与えられた時には、次回から同じような状況の下では同じ動作をしないように学習を行う。これは、まさに外界とのフィードバックループを使った自律学習である。この強化学習は、報酬や罰から学習する点が、我々生物においても観察できる非常に自然なアルゴリズムであるということができ、また、報酬や罰という非常に簡単な情報から様々な動作が学習できるという点で非常に強力な学習アルゴリズムであると言うことができる。本論文でも、この強化学習を自律学習の要として捉えていく。

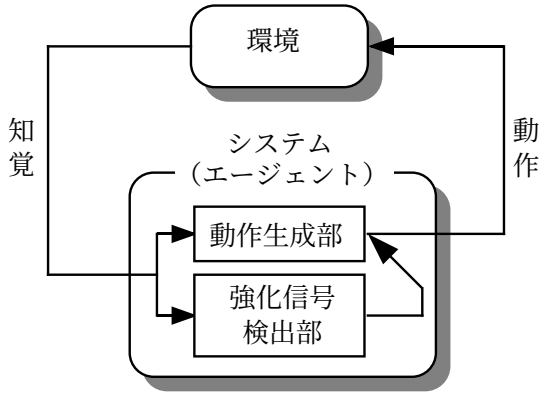

図１.６ 強化学習の仕組み

強化学習は、教師あり学習か教師なし学習かという議論がよくなされる。確かに、報酬や罰の検出方法を予め与えなければならないため、教師あり学習であると考えることができるが、出力に対し、直接教師信号を与えるわけではないので教師あり学習でもないといえる。結局、言葉の定義の問題であるが、両者の中間的存在であると考えるのが妥当であろう。

ただ、例えば、ニューラルネットにおいて教師信号と実際の出力の差の自乗を罰として与えれば、教師あり学習に近いこともできる。また、実際の生体でも何を生得的に報酬と感じ、何を罰と感じるかは判断の分かれるところである。我々がシステムに強化学習を適用する際にも、この強化信号をどう設定するかが一つの大きな問題である。このような問題に対し、何を報酬や罰にするかは進化の過程で決定されると言う考え方も出てきている。生き延びるために必要なことを報酬と感じ、死に関係することを罰と感じれば、その生物は生き延びる可能性が高くなるということは、実際の生物を考えた時にもっともで興味深い考え方である。しかし、この場合も、我々がこれを利用しようとすると、進化における淘汰の評価関数をどう設定するかという問題は依然として残る。

また、心理学の分野では、反応遮断化理論[Allison 74] [坂上 94]というものが最近受け入れられているようである。これは、強化学習を刺激-反応という図式で捉えずに、刺激も反応と置き換えることにより、強化を「複数の行動を適切な割合に再配分すること」と解釈することである。これは、ニューロンや個体は、適度な刺激と適度な（モデレートな）動作を好むというモデレーショニズム[Okabe 88]というもう一つの自律学習アルゴリズムの考え方につながるものがある（本節後述）。しかし、この場合も、やはり適度というものをどう設定するかが問題となる。いずれにせよ、実際の生物がどのようにして強化信号を決定しているか、また我々がシステムを作る際にはどうすべきかは今後のさらなる検討が必要である。本論文中では、第６章で、強化学習を認識や認識のための動作に適用することを試みたが、ここでは、認識出力がいかに理想値に近いかを強化信号として利用した。また、第７章では、強化学習の例題として、移動ロボットが目標物に到達するという問題を考えているが、ここでは、移動ロボットが目標物に到達した時に、強化信号が検出されるという設定で学習を行った。

強化学習の源は、心理学の分野でのオペラント条件付けに見ることができる。古くは、１９世紀終わり頃、 Thorndike の問題箱の実験[Hebb 72] [東 69]がある。ここでは、ネズミを箱の中に入れ、ドアに仕掛けをし、ドアの外にエサを置くことによって、試行を重ねるに従って、その仕掛けをはず

してドアを開け、エサを獲得するまでの所要時間が徐々に減少することを確認している。さらに、1961年の Skinner のスキナー箱として有名な実験がある[Skinner 61]。この実験は、図１.７のように、レバーを押すと上からエサが落ちてくる仕組みになっている箱の中にネズミを入れると、ネズミはうろうろして偶然レバーを押してエサを得ることを２・３回繰り返しているうちに、自分でレバーを押してエサを得るようになるというものである。これは、あたかもネズミがレバーを押すとエサが得られるという因果関係を把握し、さらにエサを得るためにレバーを押そうという意志を形成したかのように見える。これに対し、ネズミは単なる反射を形成しているに過ぎず、我々人間の因果関係の把握や意志とは明らかに違うと考える意見もある。しかし、所詮、我々人間も経験に基づいて、物事の因果関係を把握し、報酬を求めて意志が形成されているのであり、両者の間に本質的な違いはないと筆者は考える。このネズミの行っている学習こそが、今の知識付与型知能システムにない機能であり、自律学習（強化学習）の重要性を物語っていると考える。

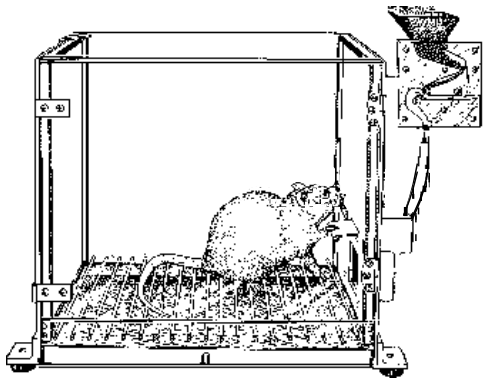

図１.７ Skinner の実験[Skinner 61]（塚田裕三編「図説 脳」日経サイエンス社より一部改変）

一方、計算論的な強化学習の研究は、1950 年の Samuel の Checker のゲームの学習に関する研究[Samuel 59] が始まりであると言われている。ここでは、ゲームの木の探索に用いる評価関数に対し、ゲームに勝ったか負けたかを評価関数として強化を行うというものである。

強化学習においては、主体が試行錯誤の成分を含む動作を行った直後に強化信号が得られれば、報酬の場合はその試行錯誤の成分を含めた動作を学習させ、罰の場合は、試行錯誤の成分と逆の成分を加えた動作を学習させることで簡単に動作を学習することができる。しかし、通常は、一連の動作の後に強化信号が得られる場合が多く、これが学習を難しくしている。通常、この問題を単に強化学習と呼ぶ。しかし、本論文では、第６章で強化学習を認識や認識のための動作の学習に適用する際に、単純化のため、毎単位時間毎に評価（強化信号）が得られると言う設定を行っている。そこで、一連の動作の後に遅れて強化信号が得られる場合を遅延強化学習、毎単位時間毎に強化信号が得られる場合を単なる強化学習と呼んで区別する。

この遅延強化学習は、Profit Sharing [Holland 87]のように、得られた報酬を使って直接過去の一連の動作を強化する方法と、報酬から中間的な評価を学習し、その評価値から動作を学習するものとに分けられる。前者は、報酬や罰が得られた時に、そこまでの動作を直接強化する方法であり、前述の

Samuel らの研究はこれに属する。前者は、直接動作を強化するため、学習は速いが、過去の全ての状態を保持しなければならない上、状態に対する評価がなされないため、汎化能力に欠けると筆者は考える。

後者の学習方法としては、1983 年の倒立振子の制御の学習を例題として扱った Barto らの critic-actor アーキテクチャ（ＴＤ学習）[Barto 83] および Watkins らのＱ学習[Watkins 92]が有名である。ＴＤ学習では、現在から将来にわたる強化信号の重み付き総和（未来ほど重みを指数関数的に小さくする）を最大化することを目標にしたものである。そして、この強化信号の重み付き総和を状態（センサ入力）から予測し、それを評価値とし、その評価値が良くなるように動作を学習するという方法である。そして、この評価値はより正しく強化信号の重み付き総和を予測するように１単位時間先の評価値の値を元に更新される。これによって、倒立振子の角度がある値より大きくなったときにペナルティを与えるだけで、倒立振子が倒れないよう制御を学習することができる（第７章参照）。

一方、Ｑ学習は、Bellman の最適化原理 [Bellman 57] に基づき、状態と動作のペアに対して評価値（Ｑ値）を設定し、そのＱ値が最大となる動作を選択するという方法である。そして、やはり１単位時間先の評価値から現在の評価値を学習する。どちらかといえば、Ｑ学習の方が、動的計画法（ＤＰ, Dynamic Programing）から直接導かれたものであり、マルコフ決定過程の上で最適な行動の獲得が保証されており、広く用いられている。しかし、元々動作は状態から状態への遷移を表していると考えれば、状態の評価が決まれば動作に対する評価も一意に決まることになる。従って、動作が評価関数のパラメータになることによって評価関数の入力空間の次元が増えることになり、汎化能力に欠ける上、連続値動作は現在のところ扱うことができないという問題点があると筆者は考える。

その後、Anderson らは、ＴＤ学習に多層ニューラルネットを用い、学習にＢＰ法を用いることによって倒立振子が倒れるまでの回数が飛躍的に伸びることを示している[Anderson 89]。また、最近は、Barto らによって、ＴＤ学習と動的計画法 (Dynamic Programing) との関係からＴＤ学習の収束性を示す議論が行われている[Barto 95-1]。また、Schultz らは、サルを用いた条件反射の学習の実験において、大脳基底核のドーパミンニューロンが、最初は報酬に反応するが、慣れると報酬に反応しなくなり、そのかわり前兆となる感覚刺激に対し反応するようになることを示している [Shultz 93]。これに対し、Barto らは、ドーパミンニューロンの反応がＴＤ学習における誤差に相当するとの考え方に基づき、大脳基底核においてＴＤ学習が行われているのではないかとのモデルが示されている[Barto 95-2][Houk 95]。

これらの動きとは別に、Widraw らは、トレーラーの車庫いれ問題を例にとり、制御対象のダイナミクスを制御信号をランダムに変化させながらニューラルネットで学習させ、このニューラルネットに制御時の誤差を逆伝播させることによって、制御信号の誤差を求め、さらにその誤差から制御信号を生成計算する別のニューラルネットの学習をさせるという学習[Widrow 85]を提案している。また、Werbos らはこの流れを汲み、システムに外界のモデルを持たせるべきであるとし、制御対象のモデルを学習させたニューラルネットと制御信号生成のニューラルネットの様々な組み合わせを提案している[Werbos 90] 。

日本では、遅延強化学習ではないが、銅谷らの歩行パターンを学習するロボットへの応用[銅谷 86]が先駆的な研究として挙げられる。ここでは、リンク機構を２つ持ったロボットにロータリーエンコーダを付け、この値を強化信号とすることによって、ロボットに歩行パターンを予め教えることなく

学習によって獲得させるている。また、依田らはヒトデの起きあがり問題への適用している[依田 90]。最近では、Onat らは、フィードフォワード型のニューラルネットの中間層の出力値を次の時間に外部からの信号と共に入力層に与える Elman 型のリカレントニューラルネット[Elman 90]をＱ学習に用い、リカレントネットに蓄えられた記憶を有効に利用して動作学習を行わせるという研究を行っており[Onat 95]、銅谷はＴＤ学習を連続時間へ拡張するという研究を行っている[銅谷 95]。

また、より実用に近い応用としては、ＴＤ学習を Backgammon というゲームに適用した例があり、人間のチャンピオンと互角の勝負をしている[Tesauro 92]。また、浅田らは、視覚センサを持ったロボットにサッカーを行わせる際に、Ｑ学習によってボールをゴールにシュートさせることを学習したり[浅田 95]、敵を避ける動作の学習とシュートの学習を別々に行った後にそれを統合するという面白い試みをしている[Asada 94]。

強化学習の解説記事として [畝見 95] [銅谷 96] が出ており、詳しくはこちらを参照されたい。

1.2.2 その他の学習指針

強化学習以外のフィードバック学習アルゴリズムとしてモデレーションニズムがある[Okabe 88][甲原 94]。これは、前述のように、各ニューロンの入出力がモデレートなレベルを好むという非常に簡単な学習則である。これによって、生体自身が適度な刺激を求めつつ過度な刺激から逃れることによって、生き延びることができるという考え方である。これは、個々のニューロンレベルで学習することを目指しており、個々のニューロンの学習を統合的に管理する部分を持つ必要がなく、フォルトトレランスの観点からも優れている。現在、人工腕での反射弓が学習によって獲得できることが示されている[浅野 95]が、今後、高次機能実現への道筋が示されることが大いに期待される。

もう一つのフィードバック学習として、筆者は、「できるだけ外界の情報を得るための動作の学習」というもの考える。例えば、我々は何か物音がすれば、そちらに目を向けるとか、動いている物体を追跡するといった動作を行う。これは、外界の情報をよりたくさん得られれば、より正しい認識ができ、より報酬を得られやすくなると考えることによって、強化学習から学習できる機能であると考えることができる。筆者も、これに近い考え方で能動認識を強化学習で学習させようとしており、第６章でこれについて述べる。しかし、外界の情報をよりたくさん得るということは、システムにとってプラスになってもマイナスになることはあまり考えられない上、これに基づく学習則は、特定の場面にしか適用できないものではなく、汎用的な学習則であると言える。また、我々が持っている好奇心というものもこれによるものと考えることも可能である。そこで、これも学習指針の一つとして挙げても良いのではないかと筆者は考える。この問題は、得られる外界の情報量を強化信号として強化学習を行っていると考えることもできる。ただし、外界の情報の量をどのように表すかは問題である。本論文中では、この学習指針はあまり多く用いていないが、第５章の物体追視のモデルにおいて、抽出した空間情報の時間変化が大きい方がたくさんの情報を得ているという観点から、より情報を得られるような動作を学習することを試みている。

以上、フィードバック学習の際の学習指針に関して考えてきたが、我々人間のようにセンサから時々刻々たくさんの情報が得られる場合には、センサの信号から前もって必要な情報を抽出し、抽出した情報から効率的にフィードバック学習を行うことも重要ではないかと考えられる。そこで、複数

種の情報源、特にセンサと運動の間や、異種センサの間に共通して存在する情報が重要な情報であるという考え方から、教師なしでその情報の抽出を学習する方法を考えた。これが、相関情報抽出学習であり、第２章でその概要を述べると共に、第３章で、詳細および空間認識との関連を探る。

さらに、センサから得られる空間情報と時間の関係が重要であると考え、その関係を学習によって獲得することを考えた。そして、センサから得られる信号を入力とし、時間と共に滑らかに変化する出力を学習する方法を提案した。これが時間軸スムージング学習であり、第２章で概要を述べる。そして、これは、センサ信号の統合および遅延強化信号に用いることができるが、これを第４章および第７章で述べる。

また、De Charms, R.C. は「全ての人間は、自己の環境に変化を生み出すことに効果的でありたいと願っており、自分の運命を自分でコントロールし、外界にもて遊ばれたくないと願っている」[Charms 76][丸野 89]という人間の行動の捉え方をしている。筆者は、 自分自身を振り返り、この考え方にも賛同する。まず、外界にもてあそばれないためには、まず外界の変化を予測することが必要である。これに関しては、前述の時間軸スムージング学習がこの役割を担うものと考えており、本論文の一つの大きな柱である。そして、外界の予測を行った上で、外界にもて遊ばれず、自分の運命を自分でコントロールするという点に関しては、強化学習がその役割を果たすと考える。また、外界に変化を生み出すことに効率的であるという点に関しては、時間軸スムージング学習と強化学習の組み合わせによって可能であると考える。

さらに人間の学習の中で、模倣というものが大きな役割を示すことも忘れてはならない。中でも、生まれて２、３カ月の子どもが、おとなが口をゆっくり開閉すると真似をしようとする [丸野 89]ことは、大変に驚くべきことである。模倣とはいえ、これを単純な教師あり学習としてすませることはできない。[丸野 89]で指摘されているように、赤ん坊がこのような機能を獲得するためには、１．口を開閉することを見ることができる、２.相手の口と自分の口が対応していることを知っている（手を動かさないで口を動かす）、３．視覚の情報と運動の情報を結び付けることができる、４．真似をすることがいいまたは真似をしなければならないと思っている、等のことが必要である。これらの全ての機能を生後（受精後？）の短期間で学習によって獲得すると考えることはかなり難しく、予め備わっている機能と考えた方が自然である。しかし、おそらく自分が口を動かすとおとなが口を動かしてくれる、または、おとなが口を動かした時に自分が口を動かすと親が非常に喜んでくれるというもっとも身近なフィードバックループが形成され、自分がおとなの動きをコントロールできるということに快感を得ているということが赤ん坊のこのような動きを学習させていると言う面もあるのではないかと考える。

また、より成長してからの模倣に関しても、最初は模倣するとまわりに喜ばれるというレベルから始まり、その後模倣することが問題解決への近道であることを学習し、さらに、模倣すること自体に快感を覚えるようになり、模倣がさらに加速され、高度な機能を身につけることに役立っていると考える。

１.３ 基本的立場

本論文の理解を深めるため、この節では、本論文に関する研究を遂行するに当たって筆者がとってきた基本的な立場を明らかする。

1.3.1 学習の段階性と並列・統合学習

学習においていきなり難しい問題を解くことは困難であるが、簡単なことから徐々に積み上げることで高度な機能の学習が可能になる。例えば、ハトを使ったスキナー箱の実験において、ハトがある場所をつつくとエサが出るという仕組みになっている場合、つつくポイントを徐々に上にもっていってやるとかなり上の方までつつくことができるようになる[東 69] 。しかし、もし、最初から高い所をつつかなければエサがでてこなければ学習不可能であろう。これは、我々人間の発達過程が、非常に無力な赤ん坊の頃から少しずつ少しずつ積み重ねていくことからも言えるであろう。例えば、赤ん坊の発達過程で、近知覚（触覚等）から遠知覚（視覚等）へと学習が進む[丸野 89]のもそのためであろう。しかし、発達の過程において、諸処の機能を、例えばある時期は歩行を学習し、ある時期は空間認識の学習をするというようにシーケンシャルに学習していくのではなく、たくさんの機能を並列に、そして徐々にその機能をアップさせているように見える。これによって、相互の機能間の関係を密接に積み上げていくことができ、柔軟かつ高度な知能に結びつくものと考える。この結果、身体の発達と機能の積み上げということから、結果的にある時期に歩行ができるようなり、ある時期にしゃべれるようになるというように学習に流れができると考える。

従来の知能システムの研究では、認識や動作のプロセスを細分化し、ある時は A という方法、ある時は B という方法をという形で、個々の場合に適した方法を考えるという形で知能化の度合いを上げていったと考えられる。一方、前節で述べたように、実際の我々は、常に様々な要因を統合して認識や動作を行っているように見える。個々の機能を個別に考える時には、そのインターフェイスを規定する必要があるが、これを予め規定してしまうことは、適応性という面から好ましくない。従って、従来のように、個々の機能を個別に考えるのではなく、できるだけ入力から出力を統一的に捉えて学習する方法が望ましい。そのためにも、上記のような段階的な学習が必要になる。

学習によらないで、ある程度発達した状態からスタートさせるためには、それだけ完成度の高いモデルを構築しなければならない。そうすると、結局与えたモデルを越えることはできないし、初めから完璧なモデルを作ることが困難であるため、システムが暴走し、人間に危害を加えることになりかねない。その点、赤ん坊のように、無力な状態からスタートすれば、いきなり暴走する心配はない。無力な状態に設定するには、赤ん坊のように身体的に無力な状態とし身体の発達と共に内部の処理も学習していく方法と、動作出力が０になるように内部の処理のパラメータを設定しておく方法が考えられる。前節で述べたフィードバック＋フィードフォワードの制御の学習の際には、後者の方法でスタートさせ、暴走することなく自然に学習を進めることができた。

制御の分野では、川人らが提案しているフィードバック誤差学習が有名である[Kawato 87]。彼が提案しているアーキテクチャは、上記の筆者が提案したものと似ているが、フィードバック制御の部分は予め備えつけられており、フィードフォワード制御をニューラルネットで学習することを試みている。ここでフィードバック制御を備え付けとしたことは、最初からある程度の機能を備え、かつ、

暴走してはいけないというという束縛によると考えられる。学習とは機能を向上させるためのものであり、赤ん坊のように機能を獲得するものという捉え方が一般的にされていないためであると考える。しかし、これからは学習の比重を増やし、機能の獲得という捉え方をするべきであると筆者は考える。

しかし、これに対し、予め備え付けられるものは備えておくべきであり、全て学習によって獲得することはナンセンスであるという反論がなされる。これはもっともなことである。しかし、予め与えたことによってそれに縛られてはならず、そこからさらに適応する機能を持たなければならないと考える。ニューラルネットで学習させることを考えれば、その初期値として情報を与えることは、そこからさらに学習させることができるという意味で有効であると思う。ただ、あらゆる場面で適応できるということは、結局、無からの学習ができる機能を持ち合わせていなければならないと考える。従って、本論文では、敢えてできる限り無からの学習をさせるということを試みた。

1.3.2　実時間学習

ニューラルネットの学習のさせ方は大きく２つに分類することができる。一つは、一般的に行われる方法で、有限のデータを何回も繰り返し学習させる方法（サンプル学習）でもう一つは、環境から逐次データを持ってくる方法である。前者は、過学習という問題を生じ、学習させるほど、また、中間層ニューロンを増やすほど汎化能力がなくなると言われている。しかし、これはある意味で当然であり、与えられたデータに対して最適にフィッティングしようとするニューラルネットの利点なのである。これに対し、実際の我々生物は、実世界との関わりを通して、常に学習を続けるため、そのような問題は起こらないし、もし、限られた範囲のデータしか取得できないような環境にいるのであれば、それに対して適応すれば十分であるということになる。従って、自律学習を目指す上で過学習を問題とすることはあまり意味がなく、実世界との関わりをしながら学習を行えば良い。また、この場合、次から次へとデータが得られるため、一部のデータを使って何回も繰り返し学習することはあまり意味がなく、１つのデータに対し１回だけ学習する。これは、汎化という点から重要であるだけでなく、リアルタイム性からも重要なことである。ただし、連続時間で考えれば、時間をさかのぼってニューロンの内部状態を元の値にセットし直して再び学習させること自体がナンセンスになるので、この問題は自ずと解決する。

しかし、通常の教師あり学習では、いくつかの入力に対して教師信号を人間が用意しなければならないため、どうしてもサンプル学習になってしまう一方、実際の環境とのインタラクションを通して学習するにせよ、環境から常に実際の確率分布に従ってデータをとってきて学習させるにせよ、存在する空間は連続であるため、無限個の入力データがあることになる。そのすべてに対して教師信号を人間が用意するのでは意味がないので、教師あり学習させるためには、その教師信号を自動生成させる手段が必要となる。

また、システムが変化する環境に適応できるようにと考えると、学習は常に続けることが好ましい。また、こうすることで、１回の学習による変化を小さくし、安定した学習ができると考える。一方、生体の発達過程を見ると、臨界期（感受性期）というものが存在し、例えば、仔ネコの片眼を遮蔽するという実験を行うと遮蔽されていなかった眼により強く反応する細胞が増えるということが報告されており、さらに、この効果は生後５週ぐらいで最大になり、それよりも早い時期でも遅い時

期でもその効果が減少することが報告されている[Hubel 72][Blackmore 76]。また、一般的にも、老化するほど学習能力は劣るという認識がある。後者に関しては、若年期にはできるだけ学習し、老年期にはそこまでに学習した結果をできるだけ活かすという学習戦略は、個体の生存を考えた時に、理にかなったものであると感じる。前者の臨界期の問題に関しては、それ以外の時期には学習が行われないというのではなく、前節で述べたような身体の発達と機能の積み上げによる学習の流れの中で、ある機能が形成できなかったため、結果的にそれ以上の積み木を積み上げることができなくなってしまったのであり、その機能にかかわること全てその時期に出来上がるわけではないと解釈することもできるであろう。以上より、筆者は、老化による学習量の減少はあるものの、学習自体は常に進行させるべきものと考えている。

本論文では、上記のような考え方に基づき、基本的にセンサを通して逐次データを得ることができるものとし、また、学習は一回のみ行うこととした。また、学習に完了はなく、常に学習をし続けるようにした。

1.3.3　遺伝（生得）と環境（学習）

心理学の分野で，人間や動物の持つ様々な行動の発達は遺伝によるものか環境によるものかという問題は，一卵性双生児の研究などを通して昔から活発に議論がなされているようである。最近では，Stern，W により提案された輻輳説，つまり，遺伝と環境の両方が影響を及ぼすものであるという説が有力である[田中 90]．筆者は、これらは複雑に絡み合っているものであり、分類すること自体にあまり意味がないと感じる。

馬などの動物は生まれるとすぐに歩くことができるが、人間が生まれた状態はあまりにも無力である。Portmann は、これを生理的早産と呼び、人間は他の高等なほ乳類と比較して約１年早く生まれていると指摘している[Protmann 44]。しかし、ほとんどの人間は大きくなれば歩くことができるようになるし、その歩き方には大きな個人差は見られない。このことから、人間の歩くという機能に関してはあまり学習の必要性がないように見える。にもかかわらず、人間が生まれた時に歩けない理由として、人間は、(1)遺伝子という限られた情報量を、学習という機能により多く費やしているため、(2)前述のように、より無からの学習をすることによって、より高度な機能を身につけることができるため等を考えている。前述の Portmann も、動物の行動は、環境に拘束され、本能によって保証されていると、我々は簡単に特徴づけることができる。これに対して、人間の行動は、世界に開かれ、そして決断の自由を持つ、と言っていいだろうと述べている。

いずれにしても、馬の歩行と人間の歩行に質的に大きな違いはなく、馬の歩行にしても、けがをしてもそれなりの歩き方ができるなど、学習によるところは大きいであろう。このことから、一見生得的な機能とそうでない機能も実はあまり大差はなく、生得的と見られる機能も、そのほとんどは実は学習によっても獲得できるものではないかと考えることができる。そして、馬が生まれつき歩けるのは、前にも述べたように、ニューラルネットの初期値が与えられていると考えることが自然ではないだろうか。

進化の研究において、個体が学習機能を身につけることで進化が大きく加速し、さらにその後、学習後の機能を予め持ったものが出現するというボールドウィン効果というものが知られている

[Baldwin 1896]。これより、学習機能がいかに重要で、進化とも切っても切れない関係であるかということがわかる。そして、人間は、まさに高度な学習機能を身につけたものであり、進化の過程の中でも、他と機能面で大きく差を付けているということができる。ただし、歩行の問題に関しては、馬が学習後の機能を予め持っているものとすれば、馬と人間の関係がボールドウィン効果の中で逆転していることになる。これはまさに前述の(1)(2)の理由によるもので、より高度な学習機能を付けるために学習後の機能を予め持つことを放棄したという逆ボールドウィン効果とも言えるものではないかと筆者は考えている。

1.3.4　生物のモデルか工学的応用か

本研究は、前にも述べたように、自律学習によって、自律的でかつ柔軟で適応性のあるシステムの構築を最終的な目標としている。自律学習のお手本は生物であり、知識付与という概念を捨てて大胆に生物に学んでいかなければならないと考える。このような意味から、本研究は、生物のモデル化を目標としているのではないが、自律学習という点からできるだけ生物から学ばなければならないと考える。ただ、心理学にしても生理学にしてもまだまだ生物のことが完全にわかっているわけではなく、未知な部分が多いし、正しいと思われていることでも研究が進むにつれて違っていたということもある。従って、各機能に関して、あまり心理学生理学に捕らわれ過ぎることがないようにした。

また、知識付与型の知能システムの場合は、その知識を付与することによって工学的にどのような機能が実現できるかということによってその善し悪しを測ることができる。しかし、本研究は、学習能力の獲得を目指し、最終的に機能実現に結びつけるというアプローチである。従って、まずはどのような機能が獲得されたかではなく、どのような学習機能が獲得されたかという観点で評価をされることが望まれる。従って、工学的応用という面からもあまり捕らわれないようにした。

筆者は、本研究が進むことによって、最終的には生物のモデル化という意味でも、工学的な応用という意味でも成果が大いに期待できるものと考えている。このような立場から、自律学習という機能にしぼって両者の中間をねらっていきたい。

１.４ 本研究の目的

本研究では、まず、自律学習という観点、つまり、いかにして自ら学習していくか、また、いかに少ない情報から多くの機能を学習によって獲得するかといった観点からシステムがどのような学習をしていくべきかを考える。この学習を、図１.８のようなシステム構成の中で、センサ信号の処理（センサ情報の統合）、および行動の生成（強化学習）にいかに展開していくかという点にある。

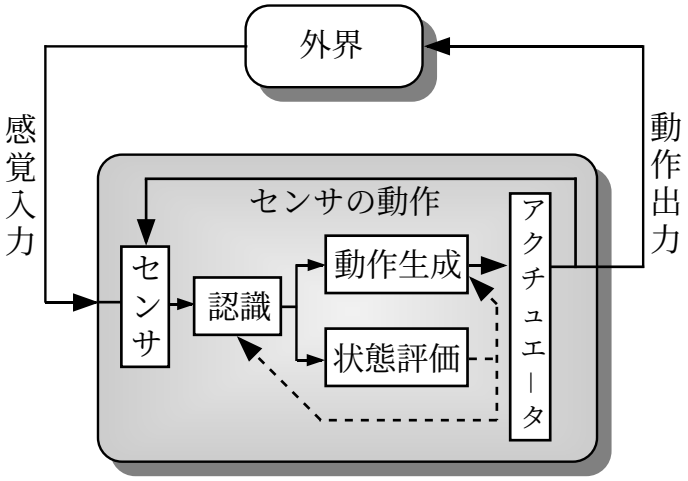

図１.８ 本研究で想定した自律学習システムの構成

１.５ 本論文の構成

本論文は、学習アルゴリズムと達成する機能の大きな２つの軸を持つ。この序論の後の第２章では、本論文中で用いる主な学習アルゴリズムについて述べる。まず、バックプロパゲーション（ＢＰ）法について述べ、その後、ＢＰ法を用いることによって簡単に強化学習を行う方法を述べる。その後、筆者が提案している、自律学習を実現する上で必要もしくは有効と考えられる３つの基本的な学習アルゴリズムである、相関情報抽出学習、時間軸スムージング学習、値域拡大学習について述べる。

第３章以降は、第２章で提案した学習アルゴリズムを用いた具体的な応用例へと展開していく。第３章では、相関情報抽出学習と値域拡大学習を組み合わせることにより、複数の情報源からの信号に共通する情報（相関情報）を抽出することを教師なしで学習する方法について述べる。そして、視覚と運動の情報を相関情報を抽出することによって、対象物体との相対位置を抽出することを学習できることを示す。また、抽出する情報が複数次元の場合に、それを直交化する方法について述べる。

第４章では、時間軸スムージング学習と値域拡大学習を組み合わせることにより、局所的な受容野を持つ多数のセンサセルよりなるセンサの信号を統合し、物体の位置などの情報をアナログ値として表現することを教師なしで学習させる方法について述べる。

第５章では、第４章で述べた学習をさらに拡張することにより、視覚システムの学習モデルとして、頭部位置によらない認識、前庭動眼反射、物体追跡の３つの機能を説明できることを示す。これによって、上記の時間軸スムージング学習を使ったセンサ信号の統合の有効性を示す。

第６章では、強化学習を能動認識、つまり、認識および認識のための動作という直接報酬や罰と関係ないものに対しても適用できることを示し、視覚センサの入力を基に認識、および視覚センサの動作を直接教師信号を与えることなく実現する方法を示した。そして、簡単な文字認識の問題を学習させることにより、センサの動作を行わせない場合より、少ない中間層ニューロン数で学習することができることを示した。

第７章では、時間軸スムージング学習の遅延強化学習への適用について述べる。ここでは、遅延

強化学習の評価を、目的達成までの所要時間で行うこととし、所要時間の予測を時間軸スムージング学習で行うというものである。また、所要時間が異なる場合も正確に所要時間を予測するために、時間軸スムージング学習を拡張した評価値の時間変化量一定化学習を提案する。以上を用いて、視覚センサ信号を直接入力としても学習できること、障害物がある場合でもある程度学習が可能であることを示すと共に、視覚センサを入力とした場合には、ニューラルネットの中間層に空間の情報を必要に応じて自己組織することを示し、その自己組織の仕方について検証した。

第８章は、結論であり、本論文の成果および今後の課題を述べる。

本論文は、表１.１に示したように、図１.８で示した自律学習システムのどの部分に適用したかと、提案したどの学習則を用いているかという２つの切り方をすることができる。そして、前者によって章の流れを構成している。第２章は、本論文で用いている学習アルゴリズム全般について概説し、その後、第３章と第４章では、センサ信号の統合・認識の問題、第５章と第６章では認識と認識のためのセンサ動作を絡ませた話し、第７章では、動作生成まで含めた話しという流れとなっている。

第３章と第４章では、（１）複数種の情報源に共通に存在する情報が重要な情報であるという観点と（２）空間情報は時間的に連続にしか変化しないため、それを用いて情報の効率的表現をすべきであるという２つの観点からその学習則を示した。また、第５章と第６章では、（１）より時間的に滑らかに変化する情報を得るように認識や認識のための動作が働くという考え方と（２）認識やセンサ動作を目的達成（強化信号を得る）ための動作として捉える、つまり、認識重点型と動作重点型の２つの考え方に基づく学習則をそれぞれ示した。どちらも同じ機能を実現するための方法が複数存在するが、実際の人間はこのどちらかというよりは、これらの複数の学習方法を組み合わせているのではないかと考えている。

表１.１　本論文における章の構成

	適用部分	用いる主な学習則	内容
第２章	３つの基本学習則の提案		
第３章	認識	相関情報抽出学習 値域拡大学習	マルチセンサまたは センサと運動の情報統合 の学習
第４章	認識	時間軸スムージング学習 値域拡大学習	局所センサ信号の統合の学習
第５章	認識＋センサ動作	時間軸スムージング学習 値域拡大学習	視覚系機能の学習モデル
第６章	認識＋センサ動作	強化学習	能動認識の学習
第７章	動作生成＋状態評価 （＋認識）	時間軸スムージング学習 強化学習	遅延強化学習