強化学習エージェントへの階層化意志決定法の導入 — 追跡問題を例に —

Reinforcement Learning Agents with Analytic Hierarchy Process: A Case Study of Pursuit Problem

片山 謙吾

Kengo Katayama

岡山理科大学 工学部 情報工学科

Department of Information and Computer Engineering, Okayama University of Science

katayama@ice.ous.ac.jp

興石 尚宏

Takahiro Koshiishi

岡山理科大学大学院 工学研究科 情報工学専攻

Graduate School of Engineering, Okayama University of Science

kohseki@ice.ous.ac.jp

成久 洋之

Hiroyuki Narihisa

岡山理科大学 工学部 情報工学科

Department of Information and Computer Engineering, Okayama University of Science

narihisa@ice.ous.ac.jp

keywords: profit sharing, analytic hierarchy process, multi-agent reinforcement learning, pursuit problem

Summary

Reinforcement Learning (RL) is a promising technique for creating agents that can be applied to real world problems. The most important features of RL are *trial-and-error search* and *delayed reward*. Thus, agents randomly act in the early learning stage. However, such random actions are impractical for real world problems.

This paper presents a novel model of RL agents. A feature of our learning agent model is to integrate the Analytic Hierarchy Process (AHP) into the standard RL agent model, which consists of three modules: state recognition, learning, and action selecting modules. In our model, the AHP module is designed with *primary knowledge* that humans intrinsically have in a process until a goal state is attained. This integration aims at increasing promising actions instead of completely random actions in the standard RL algorithms.

Profit Sharing (PS) is adopted as a RL method for our model, since PS is known to be useful even in multi-agent environments. To evaluate our approach in a multi-agent environment, we test a PS RL method with our agent model on a pursuit problem in a grid world. Computational results show that our approach outperforms the standard PS in terms of learning speed in the earlier stages of learning. We also show that the learning performance of our approach is superior at least competitive to that of the standard one in the final stages of learning.

1. は じ め に

実世界において遭遇する現実的な問題の多くは、複雑かつ動的な変化を伴う環境であるため、多くの不確実性要素を含んでいる。そのような不確実性は人間が考慮できる範疇を超越する場合が殆どであるため、現実的な問題に対して人間が設計した行動群にしたがうエージェントを予めプログラム化することには限界がある。そのような困難な問題に対しては、近年、強化学習 (Reinforcement Learning) による対応が注目されている。

強化学習 [Kaelbling 96, Sutton 98] は、予め設定された目標を達成するために自律エージェント自身が行った試行錯誤による行動に対して環境から報酬 (reward) が与えられ、報酬をもたらず行動を優先するように環境への適応を目指す学習制御の枠組みである。一般に、強化学

習における研究は「環境のクラス」と「接近の指向性」の二つの観点から分類される [山村 95]！「環境のクラス」に関しては、状態遷移がマルコフ的であるか否かで特徴付けられ、現在の状態により未来の記述が可能なマルコフ決定過程 (Markov Decision Processes, MDP) と MDP としてのモデル化が困難または有効でない非 MDP 環境が知られている。一方「接近の指向性」については、環境同定型と経験強化型が知られている。環境同定型は、結果としてなるべく大きい報酬を得るという最適性を重視する。代表例として TD 法や TD 法の発展法である Q-learning [Watkins 92] などがある。経験強化型は、学習途中でもなるべく報酬を得続けるという効率性を重視する。そのような学習法として Profit Sharing [Grefenstette 88] が代表的である。

強化学習は、不確実性を含む環境への対応が期待され

ているが、多くの学習時間（や膨大なメモリの使用）を余儀なくされる場合が多く、現在のところ、実問題への応用や実的な場面における利用において十分に対応できるとは言い難い。一般に、学習の初期では、ランダムな行動を余儀なくされることが多くあるだけでなく、設定目標へ到達した際に得られる報酬までの時間が膨大、もしくは報酬が得られず学習が進行しない場合さえある。そのような学習の状況を踏まえると、次のような応用：

- 特定の目標を達成するために迅速な対応が要求される応用（e.g., サッカーゲーム [高橋 01]）
- 安全性や確実性が求められるような応用（e.g., 交通信号制御 [参沢 00]）

では多くの問題を誘発する。これらの現実的な応用においては、試行錯誤を伴う意志決定による行動が行われる限り、迅速な対応が期待できないだけでなく、人命にかかわる可能性もある。

そのような観点から、強化学習の長所を保ちつつ、現実的により利用しやすい、強化学習をベースとする手法が要求されることは自然な流れである。このような要求に応えるためには、学習の効率化を念頭に置き、学習初期のランダムな行動を如何に抑制するかが重要な要素となりえる。言い換えると、学習初期から最善もしくはより適切な政策を高頻度を選択することに相当する。

近年、より現実的な強化学習の設計を目指した研究が行われている [Dixon 00, Unemi 00]。これらの研究は、予め学習した情報や知識を強化学習に取り入れるというものである。強化学習においては、無論、報酬を得なければ学習が進行しないので、ある程度の知識やその他の情報を利用することは、学習時間の短縮に貢献するだけでなく、本来、強化すべきではない無効ルールの発生を抑制することにもつながると期待できる。

上述の研究と類似する観点から、本研究ではエージェント自身が本来備えるべき「基礎的な知識」にもとづく意志決定の利用を考える。例えば、幼児が何らかの目標を設定し行動する際、試行錯誤的ではあるが、候補となる行動群の中から完全にランダムな意志決定・行動によってその目標を達成しようとはしないはずである。そこには、生得的もしくは本能的とも言える「基礎的な知識」などがウェイトを有する行動群として存在し、ウェイトが比較的高い行動（もしくはそれに近い行動）を選択・実行し、目標を達成すると考えられる。ここで言う「基礎的な知識」とは、エキスパートによる高度な知識を指すのではなく、ある目標を達成するために必要不可欠となるような知識を指す。

本論文では、複数のエージェントが環境との相互作用を通して、各エージェントが独立に学習するマルチエージェント強化学習に焦点をあてる。従来の強化学習に見られるような学習初期での粗雑さをできるだけ解消することに着目し、主観的な判断や曖昧さを含む評価を定量的に扱うことのできる階層化意志決定法 (Analytic Hierarchy

Process, AHP) [Saaty 80] を AHP 器として従来のエージェントへ導入した学習エージェントモデルを提案する。提案モデルにおける AHP 器は、学習エージェント自身が設定目標を達成するために本来備えておくべき基礎知識として設計され、状態認識器から与えられる情報にもとづき、より適切な行動が優先されるように候補となる行動群を重み付けする。AHP により算出される重みと学習器の重みは、ある割合で合成され、行動選択器に送られる。エージェントは、合成された重みにもとづいてルーレット選択法により行動選択し、できるだけ適切な行動を出力することで環境との相互作用を通して学習する。上述した現実的な応用では、エージェントの不完全知覚や動的な環境を想定する方が望ましいので、強化学習法としては、非 MDP 環境下でも利用でき、さらに、Q-learning よりも比較的良好な性能を示すことが知られている [荒井 98]、Profit Sharing をベースとする。Profit Sharing 強化学習エージェントに AHP を導入する本学習法の有効性を評価するために、マルチエージェント系の代表的なベンチマークである追跡問題 [Benda 85] に適用し、従来の Profit Sharing との比較を行う。その結果、提案法は、従来法に比べ、学習初期の段階から劇的に効率的な学習がなされると共に、6・4 節に示す合成比減衰法によって、基礎知識による学習への悪影響を抑制しつつ、学習後期においては従来法と同等の学習性能が発揮されることを示す。

本論文は次のように構成される。次章では、従来の強化学習エージェントモデルを示し、強化学習をある問題に適用する際に発生する問題点および提案法との関係について簡単に触れる。3 章では追跡問題の設定について記述し、4 章で Profit Sharing について述べる。5 章では AHP の概略を示し、6 章で基礎知識の定義、AHP を有する強化学習エージェントモデルおよび追跡問題に対する提案法の適用法の詳細について記述する。提案法の有効性を評価するために実験（7 章）を行い、実験結果から基礎知識の導入による学習への影響についても考察する。8 章で提案法の重要ポイントを述べ、最後に本論文をまとめる。

2. 強化学習適用の際の諸問題

強化学習における一般的なエージェントは、図 1 に示すように、センサ (Sensor) を通して得られる状態から、次の三つのモジュール、状態認識器 (State Recognition Module)、行動選択器 (Action Selecting Module) および学習器 (Learning Module) を用いることで環境との相互作用を通して学習する [山村 95]。しかしながら、ある問題に対して強化学習を適用する際には様々な困難な問題点がある。以下にその代表例を示し、提案法との関係について簡潔に述べる。

(1) 動的な環境

図 1 強化学習における従来のエージェントモデル

- (2) 不完全知覚
- (3) 報酬の分配
- (4) 探査と知識利用の間のトレードオフ

まず、最初の 2 項目、動的な環境と不完全知覚に関して簡単に説明する。マルチエージェント系においては、各エージェントが独立に学習し、それにもとづく行動によって状態遷移確率が動的に変化し、さらに遷移先も変化するという、状態遷移の不確定性の問題 [荒井 98] が生じる。一方、不完全知覚 (perceptual aliasing) の問題 [荒井 98] は、エージェントの制限された感覚入力のもとでは、異なる状態に遭遇したとしてもそれらの状態を明確に区別できず、適切な行動が不可能になることを言う。最適性を重視する環境同定型の強化学習法は、静的な環境や完全知覚を想定し適用されることが多いが、現実的な観点からは、非 MDP となる、動的な環境と不完全知覚を想定し、経験強化型の手法を利用することが有効であると考えられる。よって、本論文では強化学習法として Profit Sharing をベースとする。

マルチエージェントの強化学習における特有の事項として、3 つ目の項目である報酬の分配がある [宮崎 99a]。目標達成のために多大な貢献をしたエージェントとあまり貢献しなかったエージェントに全く同じ報酬を与えることは多くの場合適切ではないとされている。提案法では、3 章の追跡問題の問題設定にあるように、1 逃走者を 2 追跡者により捕獲するので、それら 2 追跡者は目標の達成に常に貢献する必要がある。よって、本研究では報酬の分配については考慮しないものとし、目標状態に到達した際に同値の正の報酬を各エージェントに与える (負の報酬である罰の存在は考えない)。

4 つ目の項目に関して、教師ありの学習では探査 (exploration) 中心か知識 (exploitation)*1 利用かのジレンマは起こらない [Russell 97]。教師なし学習では、そのジレンマの解決のために、 ϵ -greedy 選択等がよく用いられる。我々の方法では、基礎知識にもとづく AHP による確率的な意志決定の重み情報と学習器から得られる重み情報を合成した重みを用いて行動選択することにより学習を効率化するため、探査中心か知識利用かのジレン

*1 ここでの「知識」とはエージェントが学習してきた知識を指す。「基礎知識」とは意味が異なることに注意されたい。

図 2 追跡問題

図 3 エージェントの衝突時の対処

マがあるだけでなく、さらに「基礎知識」か「学習による知識」かのジレンマも発生する。

3. 追 跡 問 題

追跡問題 (Pursuit Problem) [Benda 85] は、マルチエージェント系におけるベンチマークの一つとして知られており、アルゴリズムの評価、エージェント間の協調などに関する多くの研究 [伊藤 01, 加藤 01, 荒井 98, 松浦 97, Ono 96, 大沢 95] の対象になっている。追跡問題には、研究者独自の設定に伴う数多くのバリエーションがあるが、本論文では、以下の設定にもとづく。

図 2(a) に示す $n \times n$ 格子状の環境を設定し、ここに複数の追跡者 (Hunter) エージェントとひとつの逃亡者 (Prey) エージェントをランダムに配置する。これを初期状態とする。各エージェントは同時に行動し、上下左右の方向に 1 コマ進むかまたは停止の行動を一つ選択する。複数のエージェントが同一の場所に存在することはないものとする。全てのエージェントが行動した単位時間を 1 ステップと呼ぶ。図 3 のように、ある行動を選んだ際にエージェント同士が衝突する場合、衝突したエージェントは 1 ステップ前の位置に戻るとする。また、格子の外枠は壁とし、各エージェントは壁を越えることはできないとする。

追跡者の視界は $m \times m$ で与え、自分の周囲 $m^2 - 1$ コマが見える。本論文では、Profit Sharing で扱う各追跡者エージェントが各ステップで観測できる状態は、追跡者の視界内で確認できる他の追跡者の場所および逃亡者の場所であり、追跡者と逃亡者を識別することが可能とする。ただし、外枠の壁は知覚しないものとする。目標状態は、図 2(b) に示すように、予め設定した複数の追跡者 (設定する追跡者の数 l は 2 とする) が逃亡者に隣接した時とする。ここで「隣接する」とは、逃亡者の上下左右のいずれかに追跡者が存在することを指す。追跡者

から見て逃亡者が斜めのマスに存在する時は隣接したとはみなさない。なお、目標を達成した際に全ての追跡者に同値の報酬 (reward) を与える。逃亡者の視覚は、追跡者の場合と同じ設定として与える。

4. Profit Sharing 強化学習

Profit Sharing [Grefenstette 88] は、遺伝的アルゴリズム (genetic algorithm, GA) を併用するクラシファイアシステムでの信用割当 (credit assignment) の方法として 1980 年代後半に提唱された。現在、Profit Sharing は GA だけではなく強化学習の枠組みにおいても利用され、さらに非 MDP となるようなマルチエージェント環境においても有用であると期待されている。

Profit Sharing (PS) は、報酬に至るまでのエピソードにおける状態 s と実際に行った行動 a の対からなるルール系列を記憶しておき、報酬が得られたときにそれまでの系列上のルールを一括して強化する学習方法である。ルール系列は次式を用いて強化する。

$$w(s_i, a_i) \leftarrow w(s_i, a_i) + f(r, i) \quad (1)$$

ここで、 $w(s_i, a_i)$ はエピソード系列上の i 番目のルールの重み、 r は報酬値、 f は強化関数である。一般に強化関数 f は、報酬を獲得した現時点からどれだけ過去であるかを引数として強化値を返す。

同一状態に対して異なる行動が選択されているとき、その間のルール系列を迂回系列と呼ぶ。一般に迂回系列上のルールを無効ルール (ineffective rule) と呼び、それ以外のルールを有効ルール (effective rule) と呼び、無効ルールは、報酬の獲得に貢献しない可能性があり、なるべくならば強化せず抑制したいと考えるべきである。我々は、政策の局所的合理性を保証する必要十分条件が証明されている [宮崎 94] の合理性定理にしたがい、無効ルールを抑制する次の強化関数を使用する。

$$f(r, j) = \frac{1}{S} f(r, j-1), \quad j = 1, \dots, W-1. \quad (2)$$

ここで、 W はエピソードの最大長、 S は報酬割引率である。なお、報酬割引率は $S \geq L+1$ とする (L は同一状態下に存在する有効ルールの最大個数である)。PS の合理性定理は、最適性を保証していないが、MDP の仮定を必要としないのでマルチエージェント系のような非 MDP 環境に対しても適用できる点に特徴がある。

PS の学習過程における行動選択法としては、ルーレット選択法が良い性能を示すことが経験的に知られている。また、ルーレット選択法は確率的に政策を自然に実現する枠組みであり、非 MDP 環境における行動選択として有効である。以上の理由から本論文では PS の行動選択としてルーレット選択法を使用する。

図 4 本の選定に関する階層構造の一例

表 1 「値段」に関する主観的判断による一対比較表の一例

値段	Book1	Book2	Book3
Book1	同じ	やや安い	非常に安い
Book2	やや高い	同じ	やや高い
Book3	非常に高い	やや安い	同じ

5. 階層化意志決定法の概略

ある問題を解決するために、我々人間が行う意志決定は主観的な観点にもとづいてなされる場合が多い。階層化意志決定法 (Analytic Hierarchy Process, AHP) [Saaty 80, 刀根 86] は、問題を解決するための代替案が幾つか与えられ、それらの代替案の中から一つを選択する際に主観的な評価に頼らざるをえない状況において利用される手法である。オペレーションズリサーチの分野等を中心として、様々なタイプの AHP やその発展法が精力的に研究されているが、本論文で扱う AHP の構造は典型的なものである。

典型的な AHP の処理の流れについて述べる。AHP では、(1) 対象とする意思決定の問題を階層構造に分解する。一般に階層構造は、問題を「目的 (goal)」「評価基準 (criteria)」「代替案 (alternatives)」の関係で捉えることで構築される。問題の「目的」と「代替案」は予め与えられる。「評価基準」は各要素間の一対比較の際に相対的な重み付けを行うための評価の基準であり、より複雑な評価基準を複数の階層を用いることで構築する場合もある。(2) 各階層の要素間の一対比較にもとづき一対比較表を作成し重み付けを行う。それらの重み付けにもとづき、(3) 階層全体の重み付けを行うことで、各代替案の重要度を算出し、「目的」に対する代替案の優先度を定量的に決定する。

例として、三つの本 (Book1, Book2, Book3) の中からどの本を買うべきかを選定する問題について考える。この例の場合、目的はどの本を買うかという「本の選定」となり、代替案は三つの本である「Book1」「Book2」「Book3」が与えられる。評価基準として「値段」や「文学性」「思想性」を挙げたとすると、その階層構造は図 4 のようになる。

例えば、人間の主観的判断により「値段」の評価基準で各本を評価した場合に表 1 に示す結果が得られたとす

表 2 AHP における一対比較値

値	1	3	5	7	9
定義	同程度に重要	やや重要	かなり重要	非常に重要	絶對的に重要

表 3 「値段」に関する一対比較表の一例

値段	Book1	Book2	Book3	幾何平均 GM	重み w
Book1	1	3	7	2.759	0.6908
Book2	1/3	1	1/3	0.481	0.1204
Book3	1/7	3	1	0.754	0.1888

る．ここでは「安い」という評価が評価値としては高く、「高い」という評価が評価値としては低いとする．しかしながら、このような曖昧な表現では各本の評価を数量化することは困難である．そこで AHP では、曖昧な表現を表 2 に示すような整数値（一対比較値）に置き換える．表 1 の「やや安い」の場合は 3、「非常に安い」の場合は 7、というように置き換える．これにより、表 1 は表 3 のようになる．同様に、全ての評価基準にもとづいてそれぞれの一対比較表を作成する（本例では「値段」の他に「文学性」「思想性」の一対比較表ができる）．さらに評価基準（「値段」「文学性」「思想性」）の間の一対比較も行う．この場合は 1 つ上の階層の要素（本例では「本の選定」）に関して一対比較表を作成する．

一対比較表の特徴は 2 つある．1 つ目は対角要素を 1 にすること、2 つ目は、要素 i からみた要素 j の一対比較値を v_{ij} とする場合、その対角要素の値は $1/v_{ij}$ とすることである．よって、一対比較表の対角線より上の評価を定めることによって対角線より下の評価が可能になり、一対比較表が作成される．

作成された一対比較表から各項目の重みを計算する方法として幾何平均法が良く知られている．幾何平均法は各項目の行を幾何平均し、重みは幾何平均値の合計が 1 になるように正規化することで得られる．つまり、一対比較表の i 行目の幾何平均値 GM_i を $\sqrt[k]{v_{i1} \cdot v_{i2} \cdots v_{ik}}$ により算出する．ここで、 v_{ij} ($i, j \in \{1, \dots, k\}$) は i 行 j 列にある各要素の比較値である．各幾何平均値の和 $sum = \sum_{i=1}^k GM_i$ を求め、 i 行目の重み $w_i = GM_i / sum$ をそれぞれ算出する．

表 3 の場合、各行の幾何平均は Book1 : 2.76, Book2 : 0.48, Book3 : 0.75 となり、幾何平均の和は 3.99 となる．この和で各行の幾何平均値を割ると Book1 : 0.69, Book2 : 0.12, Book3 : 0.19 が求められ、値段の場合では Book1 が最も好ましいことが示される．上述したように、この操作を全ての評価基準について行い、さらにどの評価項目が重視されるかの判断も同様の操作を行うことで導き、最終的にどの代替案を選定すべきかの判断を下す．

最終的な代替案の重みは、代替案が有する重みと目的に対する各評価項目の重みを掛け合わせ、それらを加算することで求められる．以下では、Book1 を例にとり、

表 4 目的および各評価項目に対する一対比較表

(a) 目的に対する一対比較表

目的	重み w
値段	0.1428
文学性	0.4286
思想性	0.4286

(b) 各評価項目に対する一対比較表

	値段	文学性	思想性
重み w			
Book1	0.6908	0.0976	0.1350
Book2	0.1204	0.3879	0.2808
Book3	0.1888	0.5145	0.5842

表 5 各代替案の最終的な重み

	重み w
Book1	0.1924
Book2	0.3264
Book3	0.4812

問題に対する最終的な重みの算出手順を示す．各項目ごとに判断を行った結果、表 4 が得られたとする．そこで表 4 の (a) から目的に対する重みと、(b) の表から Book1 の各評価項目に対する重みを用いると、 $0.1428 \times 0.6908 + 0.4286 \times 0.0976 + 0.4286 \times 0.1350 = 0.1923$ となり、この値が Book1 の最終的な重みとなる．このような処理を Book2, Book3 とともにを行うと表 5 が得られる．よって、表 5 の結果から Book3 を選択することが適切であると判断を下させる．このように AHP では、主観的な判断にもとづき、全ての代替案に対して重視すべきその比重を数量化できる．従って、例えば Book3 が在庫切れ等であれば、次に重みの高い Book2 を選択することが適切であると判断できる．このような処理をエージェントに組み込むことで、エージェントは各状況に応じてより適切となり得る判断・行動を可能にすると期待できる．

AHP では、非整合度尺度 (inconsistency ratio, IR) と呼ばれる指標がある．これは、主観的判断による一対比較が適切になされたかを示すためのものである．本論文では、エージェントに導入する AHP において IR の整合性のチェックは行わないものとする．その理由は 6.3.3 節で述べる．

6. 提 案 法

本章は、従来の強化学習エージェントに対して基礎知識にもとづく AHP 器を導入した学習エージェントモデルを示すと共にそのモデルを用いた学習の方法論を示す．以下では、基礎知識、AHP 器を導入したエージェントモデル、追跡問題に対する基礎知識にもとづく AHP の設計、AHP 器と学習器の重み合成の順に記述する．

6.1 既知の目標と基礎知識

強化学習は、目標状態に到達した際に与えられる報酬を手掛かりにエージェントが学習する学習法であるため、報酬が得られなければ学習は進行しない．従って、報酬を

確実に得るための工夫を考える必要がある．そのような工夫の一つとして，副目標に対する報酬や制約違反に対する罰を導入することがあげられる．しかしながら，副目標の導入等が学習自体に副作用を及ぼしたり，悪影響を与える場合もある [宮崎 99b]．そこで我々は，人間により予め設定される目標状態に対する（後述する）知識の導入を考える．これにより，最初の報酬を得るまでに必要となる試行錯誤の行動の増大をある程度抑制することができると共に，複雑化する問題への対応もより現実的なものになると考えられる．なお，提案法を適用できる問題は，最終の目標（または副目標）に対する基礎知識が存在し，その設定が可能であるような問題である．

ある問題へ強化学習の適用を考える場合，その環境が全く未知であることは希である．また，強化学習において，エージェントに行わせたい目標は，我々人間にとっては既知であり，エージェントに期待する未来の結果である．そのような観点から，ある問題における目標の存在は人間が設定する限り既知であるから，エージェントに導入される知識はその既知である目標に対する基礎的なものと捉えることができる．我々はこれを「基礎知識」と定義する．

例えば，3 章で述べた追跡問題で設定される目標状態は「複数の追跡者が逃亡者に隣接すること」であるから，1 追跡エージェントにとって明らかに必要な基礎知識は，上下左右等に移動できることをはじめ，「追跡者は逃走者に接近するように行動する」ことである．他例としてサッカーゲームの場合，ゴール前にいる攻撃側のエージェントの基礎知識は，「できるだけゴールに向けてボールを蹴るように動作する」ことであり，ゴールのない方向にボールを蹴ることではない．

そのような知識を如何にしてエージェントに導入し，優れたエージェントを実現するかが問題となる．本論文では，各状態においてエージェントができるだけ適切な意志決定にもとづき行動するために必要となる基礎知識を AHP によって設計し，AHP 器をエージェントへ取り入れることで実現する．

6.2 AHP を用いるエージェントモデル

AHP 器を導入したエージェントのモデルを図 5 に示す．その構成は図 1 に示した従来のエージェントモデルの学習器に AHP 器を併用するように加えたものである．本モデルでは，学習器と同様に，状態認識器から得られる状態にもとづいて，各行動の重み付けを AHP 器で行い，行動選択器ではその重みに従ってルーレット選択法により行動選択する，という新たな流れが加わっている．よって，AHP 器への経路を除くことで従来のエージェントモデルとなる．

AHP 器は状態認識器から与えられる情報にもとづき，より適切な行動が優先されるように候補となる行動群を重み付けする．本モデルにおいてエージェントは，AHP

図 5 AHP を有するエージェントモデル

図 6 追跡問題に対する AHP の階層構造

器と学習器の重みを合成し，その合成された重みにもとづいて行動選択する．その詳細については 6.4 節で述べる．

重要な事項として，提案法では状態認識器からの状態と実行された行動の対のルールは常に学習器に送られる．よって，例え AHP 器の重みのみが利用される場合であってもエージェントは学習することに注意されたい．

6.3 AHP 器の設計

ここでは，追跡問題を例にとり，AHP 器の設計において必要となる事項について記述する．

§ 1 階層構造

追跡問題での目標は，複数の追跡エージェントが 1 逃亡エージェントに隣接することである．したがって，隣接するために必要となるシングル追跡エージェントの基礎知識は「逃亡者に近づくこと」であることは自明である．その基礎知識にもとづくと，AHP の階層構造は，目的が「逃亡者に近づくこと」，評価基準が「逃亡者の位置」，代替案が「上」「下」「左」「右」と「停止」からなり，図 6 のようになる．基礎知識に関する注意事項として「逃亡者を壁に追いつめること」や「他の追跡エージェントが待ち伏せすること」などの知識は，上述の基礎知識に比べ発展的であり，しかも協調的な要素が強くなるため，ここではシングルエージェントに与える基礎知識としては不適切とする．これらのプログラム化しにくい要素は，マルチエージェントの学習の段階で，その実現を期待するものとする．

§ 2 エージェントの視野設定

追跡者が「逃亡者に近づくこと」として必要な要件は，逃亡者の位置に関する情報である．追跡者は，あるステッ

図 7 追跡問題におけるエージェントの視野

ブにおいて視野内に逃亡者がいる場合、逃亡者の位置情報を環境の状態として有している（3 章参照）。AHP 器は、追跡者を中心とする視野内において逃亡者がどこにいるか（上、下、左、右、右上、右下、左上、左下）^{*2}に応じ、代替案である追跡者の行動（上下左右と停止）を評価する。図 7 はエージェントの視野に関する図であり、追跡者から見て逃亡者が左上にいる場合の例である。

§ 3 代替案の重み付けのための段階的な更新規定

上述した本の選定の例では、表 2 に示す一対比較値から人間の主観的判断に応じて値を選択し、一対比較表を作成したが、提案法ではその判断をシンプルに捉え、その作成を「段階的」な更新規定により自動化する。

各エピソードの初期設定として、一対比較表の全ての値を表 2 に示した「同程度重要」の 1 にセットする（表 6）。その後の各ステップでは一対比較表は次のように更新される（増減の方針については 6.3.4 節を参照）。エピソードのあるステップにおいて、例えば、追跡者から見た逃亡者の位置が右上である場合、現在の一対比較表で蓄えられている一対比較値を段階的に更新することにより代替案の右と上の重みの情報が増加するようにする。ここで「段階的」とは、表 2 に示した整数値にしたがい、エピソードの t 番目のステップにおける一対比較表で該当する一対比較値（評価値）が 1 であった場合、 $t+1$ 番目（現在）のステップではその値を 1 段階上げて 3 に更新することを指す。また、評価を下げる場合も同様に、ある一対比較値が 9 である場合は、1 段階下げて 7 に更新する。なお、一対比較表で利用される上限の値は表 2 にしたがって 9 とし、下限値はその 9 の逆数 $1/9$ とする^{*3}。このような数値の範囲を採ることで、一対比較表から算出される重みの秩序性が保たれる。表 7 は、追跡者から見た逃亡者の位置が右上であった場合に、表 6 から新たに更新される一対比較表である。

このように、AHP 器では、エージェントが観測する環

表 6 一対比較表の初期設定

位置	上	下	左	右	停止	幾何平均	重み
上	1	1	1	1	1	1	0.2
下	1	1	1	1	1	1	0.2
左	1	1	1	1	1	1	0.2
右	1	1	1	1	1	1	0.2
停止	1	1	1	1	1	1	0.2

表 7 逃亡者が右上にいる場合の一対比較表の例

位置	上	下	左	右	停止	幾何平均	重み
上	1	3	3	1	3	1.933	0.3333
下	1/3	1	1	1/3	1	0.644	0.1111
左	1/3	1	1	1/3	1	0.644	0.1111
右	1	3	3	1	3	1.933	0.3333
停止	1/3	1	1	1/3	1	0.644	0.1111

境の各状態にもとづいて、所定の規定の下で一対比較表の値が段階的に更新され、代替案の重みを算出する。従って、AHP 器で行われる評価は、従来の AHP の評価に用いられるような、人間による主観的判断ではないため、5 章で触れた IR の整合性のチェックは意味をなさないことになる。

§ 4 行動評価の増減の方針

上述したように、一対比較表は各状態に応じて自動的に更新される。その自動的な更新に伴い、想定される状況の変化パターンに応じて、追跡者の各行動評価の増減の方針を決定する必要がある。その方針を次に示す。

- [エピソードの初期状態] では二つの場合分けがある。
 - 視野内に逃亡者がいる場合は、逃亡者に近づく行動の評価を 1 段階上げ、近づく行動の評価は 1 段階下げる。関係の無い行動の評価は変更をしない。
 - 視野内に逃亡者がいない場合は、判断がつかないので各行動の評価は変更しない。
- [エピソードの途中で逃亡者が見える状態から見えない状態] になった場合は、今までに蓄積されてきた一対比較表が適切か否かの判断がつきづらくなるため、高い評価（「同程度重要」よりも高い評価を指す）は 1 段階下げ、低い評価（「同程度重要」よりも低い評価を指す）は 1 段階上げることで、各行動の評価を「同程度重要」（表 2 参照）の一対比較値（1）に近づけるようにする。

AHP 器での代替案の重みは、基礎知識にもとづくエージェントの行動の重みの量として扱うことができ、エピソードの各ステップにおいてエージェントが観測する状態によって刻々と変化する。よって、AHP 器から得られる各行動の重みの役割は学習器の場合と似ている。このことから、エージェントは、例え学習器からの重みを利用しない場合であっても、AHP 器で得られる代替案の各行動の重みを用いてルーレット選択法により行動の選択も可能になる。

提案モデルでは、あくまでも、人間が与えた基礎知識にしたがい階層構造化された AHP 器により、エージェ

^{*2} 例えば、逃亡者の位置が「上」であるとは、追跡者から見て、真上に逃亡者が存在することを指す。他の下、左、右も同様に考える。また中間的な方向として、例えば「右上」の場合、真上と真右の間の比較的広い領域に逃亡者が存在することを指す。その他の右下、左上、左下も同様に考える。

^{*3} 上限および下限の値は、それぞれ 9 や $1/9$ ではなくとも、その範囲を拡大しても良い。

ント自身が確率的な意志決定のもとで行動できる．よって、AHP 器による確率的な意志決定は、その行動自体が常に支配されるのではなく、人間が期待しなかった振る舞いを実現する可能性が残されている．例えば、表 7 を例にとると、逃亡者が右上にいる時でさえ、各行動の重みは「上」「右」だけが与えられるのではなく、右と上以外の行動も重みとして与えられるため、ルーレット選択法により、右と上以外の行動が選択されることもある．このことから AHP 器は、様々な状況に応じてより適切となり得る行動（追跡問題の場合は近づく）を選択されやすくし、適切でない可能性が高い行動（追跡問題の場合は遠のく）が選択されにくくなるように機能する．

6.4 AHP 器と学習器の重み合成

AHP 器が担う役割は学習器に似ている．図 5 に示したように、AHP 器は学習器と同様に状態観測器から得た状態にもとづき各行動の重みを算出し行動選択器に受け渡す．これにより、学習初期において良いとは言えない学習器の性能を補うことができると考えられ、AHP 器からの重みを用いることで、各状態ごとに適切な行動を選択しやすくできる．

以下では、AHP 器と学習器の各行動の重みを合成する方法について記述する．両器は併用されるので、学習途中で最も良い切り換え時期を模索するののも一つの手ではあるが、ここでは、次式のように、合成比 $rate$ を用いて AHP 器と学習器から得られる重みを合成する．

$$TWs = rate \cdot AHPWs + (1 - rate) \cdot LMWs \\ (0 \leq rate \leq 1)$$

ここで、 TWs は最終的に行動選択器に送られる各行動の重みであり、 $AHPWs$ は AHP 器から得られる各行動の重み、 $LMWs$ は学習器からの各行動の重みである．ただし、合成の計算を行う前に $AHPWs$ および $LMWs$ とともに行動群の合計がそれぞれ 1 になるようにする．本論文では、上記の合成式にもとづき固定された $rate$ を利用する方法を「合成比固定法」と呼ぶ．

学習の進行過程において、予め与えた固定の合成比を用いて AHP 器の利用を続けることもできるが、その場合、AHP 器は学習器のように長期的な改良を行わないため、学習後期においてエージェントが望ましい行動を最終的に獲得できなくなる恐れがある．そこで我々は、学習の初期段階では、良い性能とは言えない学習器の重みの利用を控え、AHP 器の重みを重視するようにする．そして学習が進むにつれて、AHP 器の重みの利用を徐々に減衰させることにより、最終的には学習器のみの重みを利用する．本論文では、この方法を「合成比減衰法」と呼び、次式によって合成比 $rate$ を減衰させる．

$$rate = \alpha \cdot rate \quad (0 \leq \alpha \leq 1)$$

ここで α は減衰率であり、減衰は 1 エピソードごとに行

う．なお、合成比 $rate$ と減衰率 α の設定値により様々な学習のバリエーションが可能となる．

7. 実験結果

本章では、学習初期および後期における提案法の有効性を評価すると共に、得られた実験結果から、基礎知識の導入による学習への影響についても考える．ここでは、3 章で記述した追跡問題の基本設定にしたがい、異なる格子サイズ ($n = 7$ と $n = 15$) の問題に対して逃亡者の行動が逃避的か非逃避的かの計 4 つの Case について実験を行う．ここで、逃亡者が「非逃避的」とは、追跡者の位置に関わらずランダムに行動することである．一方、「逃避的」とは逃亡者の視野内にある最も近い追跡者から遠ざかるように行動することであり、逃亡者の視野内に追跡者がいない場合はランダムに行動する．4 つの Case は次のようになる．

- Case1 格子サイズ: 7, 逃亡者: 非逃避的 (ランダム)
- Case2 格子サイズ: 7, 逃亡者: 逃避的
- Case3 格子サイズ: 15, 逃亡者: 非逃避的 (ランダム)
- Case4 格子サイズ: 15, 逃亡者: 逃避的

なお、問題の難しさの点からは、Case1 が最も簡単であり、Case4 が最も困難である^{*4}．

2 追跡者および 1 逃亡者の各エージェントの視覚サイズは全ての Case について 7×7 ($m = 7$) とし、Profit Sharing 強化学習の初期ルールを重みを 0.1、報酬値を 1.0、公比 0.2 の等比減少関数を強化関数とする．6.4 節で述べたように、合成比 $rate$ と減衰率 α の設定値により様々な学習のバリエーションが可能である．従来のエージェントモデルを有する Profit Sharing は $rate = 0$ および $\alpha = 0$ とすることで実現できる．また、AHP 器からの重みのみを利用（つまり、全エピソードにおいて学習による知識を全く利用せず、基礎知識のみを利用）する場合は、 $rate = 1$ および $\alpha = 1$ とすることで可能である．本実験では、従来法 ($rate = 0, \alpha = 0$) も含め、その他、減衰率 α を 1 に固定し、合成比 $rate = \{0.25, 0.5, 0.75\}$ をそれぞれ初期設定として与える場合と、初期の合成比 $rate$ を 1 とする合成比減衰法の減衰率 $\alpha = \{0.99, 0.999, 0.9999, 0.99999\}$ をそれぞれ与える場合、さらに、AHP 器のみを利用する場合 ($rate = 1, \alpha = 1$) も加えた計 9 通りの学習について検討する．各 Case に対する各学習の試行回数は 10 回である．1 試行の学習は 10 万エピソードまでとする．

なお、上記した減衰率 α の違いによる合成比減衰の傾向は図 8 のようになる．例えば (c) の減衰率 $\alpha = 0.9999$

^{*4} Case1 ~ Case4 に関して文献 [荒井 98] で示された「追跡問題の分類」に当てはめると、Case1 と Case3 が [荒井 98] での「ケース 4」、Case2 と Case4 が「ケース 5」に相当し、いずれも状態遷移の不確定性および不完全知覚の問題の点で困難な設定となっている．なお、[荒井 98] ではトラス環境を設定している．

図 8 α の設定値による合成比減衰の傾向の違い

の曲線では、AHP 器の重みを利用する割合が約 6 万エピソード付近へ向けて徐々に低くなり、それに反して学習器の重みを利用する割合が徐々に増加する。それ以降は学習器の重みだけが利用され、従来の強化学習アルゴリズムと同等の処理になる。

7.1 学習初期の結果

図 9 から図 12 までの各図は、各 Case に対する学習初期（2500 エピソードまで）の学習結果である。なお、縦軸は 1 エピソードごとに目標を達成するまでに要した行動（ステップ）回数、横軸はエピソード数である。図 9 から図 11 での線 (A) は従来法 ($rate = 0, \alpha = 0$) の結果を示しており、(B) は $rate = 0.5, \alpha = 1$ の際の合成比固定法、(C) は合成比減衰法により算出された最良の結果 (α の値については線 (C) の説明欄を参照)、(D) は AHP 器のみを利用する場合 ($rate = 1, \alpha = 1$) の結果を示している。図 12 の (B) ~ (D) については上述の通りであるが、従来法の結果 (A) は計算機のメモリ不足により実行不能となったため示していない。

各図から、AHP 器を用いる本エージェントは従来法の学習初期において多く発生する無作為な行動を排除することが可能であることを確認できる。さらに、初期の合成比 $rate$ を 1 とする合成比減衰法にもとづく学習と $rate = 0.5, \alpha = 1$ とする合成比固定法の学習の両傾向から、学習初期に AHP 器を多用する方がより顕著にその効果を発揮することが観測できる。

しかしながら、学習初期においてさえ、AHP 器を多用することは必ずしも最善ではないことがうかがえる。これは、Case1 (図 9) および Case3 (図 11) における約 1000 エピソード以降の時点から観測でき、AHP 器のみ（つまり、基礎知識のみ）を利用し続ける非常に単純な方法 ($rate = 1, \alpha = 1$) の場合では、提案法だけでなく従来法よりも多くの行動（ステップ）回数を必要とすることがわかる。よって、Case1 および Case3 に対するこの観測結果から「強化学習に導入される基礎知識が有効に働くのは学習の比較的早い段階である場合が多く、基礎

知識を利用し続けることは強化学習に悪影響を与える」ことを示唆している。

一方、Case2 (図 10) と Case4 (図 12) については、合成比減衰法にもとづく学習法と基礎知識のみを利用する方法の結果に大きな差は（2500 エピソードまでの段階では）見られなかった。しかしながら、上述の Case1 と Case3 での結果を踏まえると、より困難な問題である Case2 と Case4 における 2500 エピソードまでの学習期間は強化学習の効果を顕著に発揮できない比較的早い段階（または基礎知識が有効に働く段階とも捉えられる）であると予測されるため、学習がさらに進行した学習後期の結果からその差を確認できるかも知れない。

7.2 学習後期の結果

表 8 は、上述の計 9 通りの各学習に対して、10 万エピソードの学習を行った後に 100 エピソード分の行動（ステップ）回数を計測し、その行動回数の平均とその標準偏差の結果である。Case4 は、他の Case に比べて困難な問題であるため、従来法による学習 ($rate = 0, \alpha = 0$) は非常に膨大なステップ数を余儀なくされ上述の通り実行不能となった（表中では“—”と記した）。一方、本エージェントはそのような困難な問題に対しても実行可能であった。

表 8 の結果から、AHP 器を使用する割合を等比的に減少させることによって、従来のエージェントの場合と遜色ない（もしくは良好な）性能が学習後期に得られることを確認できる。特に、減衰率を 0.999 や 0.9999 とする多くの場合、従来法に比べてより少ないステップ数で目標状態に到達している。つまり、従来法では Case1 で 9.7、Case2 で 23.3、Case3 で 135.7、Case4 で実行不能であったのに対して、提案法の合成比減衰法の場合では、それぞれの Case に対して、9.1、20.6、128.1、644.9^{*5}であった（表中の Avg. の太字参照）。

また、減衰率 α の設定により学習性能に影響があることもわかる。例えば、Case1 では 0.9999 や 0.99999 を用いるよりも 0.999 の方がより良い結果を算出している。これは、比較的簡単な問題である Case1 に対しては AHP 器を長時間用いるよりも早々に学習器の重みの比率を高くする方が良いことを示している。Case3 では 0.99 や 0.999 よりも 0.9999 と設定することで良好な結果が得られているので、より困難な問題の場合には基礎知識をより長期にわたり徐々に低減させた方が良いことを示唆している。上述の両 Case と同様の傾向は Case2 と Case4 にも見られるので、より良好な学習性能を得るために設定すべき減衰率の値は、問題の困難さに依存するものと考えられる（なお、強化学習を最終的に打ち切る学習

*5 Case4 の問題に対して学習回数を 100 万エピソードまで増やし同様の実験を行うと、平均行動（ステップ）数が 644.9 から 444.0 に改善された。これより、表 8 の Case4 の結果は学習途中の結果であることに注意されたい。

図 9 Case1 に対する学習初期の結果

図 11 Case3 に対する学習初期の結果

図 10 Case2 に対する学習初期の結果

図 12 Case4 に対する学習初期の結果

回数にも関連すると推測される)。一般に、対象とする問題の困難さをあらかじめ予測し、減衰率や学習の打ち切り回数を設定することは難しい。一つのより望ましい解決策として、学習の状況等に応じて AHP 器と学習器の重みの合成比を適切に決定可能とする適応型の合成法が考えられる。これについては今後の課題とする。

次いで、合成比固定法の結果について考える。表 8 の結果から、合成比固定法により最終的に得られる学習性能は、合成比減衰法だけでなく従来法よりも多くの場合に劣っている。この理由として、合成比固定法は、学習初期の段階では基礎知識の利用によって従来法よりも良好な結果を示すことが可能であったが、学習後期では、基礎知識を長期にわたり固定して利用するため、学習自体に悪影響を与えたことが原因と考えられる。例えば、Case1 に対して、合成比固定法の $rate$ の設定値で比較を行うと、0.25 と設定した場合が最も良く、 $rate=0.75$ とした場合が最も悪い結果となっている。この結果は、基礎知識を用いる割合が多い場合には、学習への悪影響が特に強くなることを示す例であると考えられる。この例の極

端なものが、次に比較する、AHP 器のみを利用する方法 ($rate = 1, \alpha = 1$) による結果である。

学習初期の結果である Case2 (図 10) および Case4 (図 12) においては、合成比減衰法と AHP 器のみを利用する方法 ($rate = 1, \alpha = 1$) との比較より、ステップ数に関して大きな差はなかったものの、表 8 の結果はその差を如実に表している。例えば、Case2 では、AHP 器のみを利用する場合はその回数が約 112 であったのに対して、合成比減衰法の $rate = 1, \alpha = 0.9999$ では約 21 であった。特に、本論文で対象とした追跡問題の場合は、最終目標と基礎知識の設定が直接関連するものとなっているので、一見、基礎知識のみを利用する方法と提案法 (特に、合成比減衰法) の間の結果に大きな差は生じないと推測できるかも知れない。しかしながら、表 8 の結果から、その差は顕著に表れ、上述した合成比固定法の結果も踏まえると (少なくとも追跡問題の場合) 基礎知識を長期間固定的に利用することは学習に対して悪影響を与えることが確認された。よって、基礎知識を持たない従来法と同等 (もしくはそれ以上) の結果を得ることが

表 8 学習後期の結果—AHP 器の割合 $rate$ と減衰率 α の影響

$rate$	α	Case1		Case2		Case3		Case4	
		Avg.	Stdv.	Avg.	Stdv.	Avg.	Stdv.	Avg.	Stdv.
0	0	9.7	3.7	23.3	22.7	135.7	59.6	—	—
0.25	1	13.7	3.7	26.4	10.1	137.9	45.2	1122.9	732.2
0.5	1	21.8	6.4	37.1	9.5	151.7	57.5	840.9	221.6
0.75	1	38.4	11.3	47.5	17.1	197.7	88.8	886.3	322.5
1	0.99	9.4	3.4	20.9	12.7	145.1	50.7	718.3	405.4
1	0.999	9.1	3.1	20.6	10.9	138.8	51.5	652.0	400.5
1	0.9999	9.3	3.2	21.5	10.3	128.1	42.9	644.9	391.8
1	0.99999	18.4	4.7	29.2	12.0	177.0	53.3	832.7	425.2
1	1	87.4	25.0	112.2	33.8	936.1	292.6	1262.6	756.8

できる合成比減衰法は、基礎知識の導入による学習への悪影響を抑制できる有用な方法であると言える。

以上より、AHP を併用する提案法（特に、合成比減衰法）は、学習初期の段階から劇的に効率的な学習がなされると共に、基礎知識による学習への悪影響を抑制しつつ、学習後期においては従来法と少なくとも同等の学習性能が発揮されることを確認した。

8. ディスカッション

ここでは、提案法の重要なポイントについて論考する。

- (1) 提案法は一般化された強化学習の枠組みである。
- (2) 複数の基礎知識を導入することや各エージェントの役割に応じて AHP 器を設計することも可能である。
- (3) AHP 器での時間量は多項式時間である。
- (4) AHP 器から得られる各行動の重みには多様性がある。
- (5) AHP 器は短期記憶として機能するため、エージェントが不知覚の状態になっても、適切な行動が選択される傾向をある程度維持することができる。
- (6) 合成比減衰法により、基礎知識の悪影響による学習性能の低減が可能である。

ポイント (1) に関して：提案法は、AHP 器および学習器をほぼ独立的に利用でき、強化学習を内包する枠組みとなっているため、強化学習の一般化として捉えることができる。また、学習器と AHP 器はほぼ独立に設計可能であるため、他の（より高性能な）強化学習法を学習器に導入することができる。

ポイント (2) に関して：本論文では追跡問題に対しては基礎知識を一つだけ想定したが、実用的な問題では複数のタスクが存在し複数の知識を必要とする場合がある。AHP の考え方自体は様々な意志決定の場面で利用できるので、提案したエージェントモデルの AHP 器では複数の基礎知識を設計することも可能である。加えて、本論文では一つの基礎知識に対して一つの評価基準を階層構造として設定したが、AHP では複数の評価基準を採用することが可能である（むしろ AHP では複数の評価基準の存在が本来自然である）。また、非等質のエージェント

の設計が必要な場合、強化学習システム自体に複数の AHP 器を利用することができ、エージェントの役割等に応じて特有の基礎知識を与えることも可能である。

ポイント (3) に関して：実際の利用では、強化学習の処理時間も含め、AHP の処理に多くの時間を費やしてはならない状況も考えられる。例えば、ロボカップサッカーのシミュレーション部門で利用されるサッカーサーバでは、100ms を単位として試合の中断なしに実行されなければならない [高橋 01]。

AHP 器においては、ポイント (2) の内容も含め対象とする問題によって必要な処理時間は異なると考えられるが、一般的に AHP での時間量は一対比較表からの重みの算出に必要な時間で議論でき、多項式時間である。例えば、本論文で対象とした追跡問題の場合、代替案（行動）の数を N_a 、評価基準数を N_c とすると、各ステップで AHP の処理に要する時間量は、 $N_a > N_c$ の時は $O(N_a^2)$ であり、 $N_a < N_c$ の時は $O(N_c^2)$ である。

ポイント (4) に関して：AHP 器は、状態認識器から与えられる情報にもとづき、より適切となり得る行動が優先されるように候補となる行動群を重み付けする。ルーレット選択法のような確率的な選択法を行動選択器で利用する場合、AHP 器から算出される重みは各行動に対してそれぞれ与えられるため、例え AHP 器からの重みのみを利用していても、選択される行動には多様性があり、特定の行動だけが支配的に実行されるわけではない。

ポイント (5) に関して：宮崎らは、学習エージェントにとって「何も見えない」状態での行動の学習が非常に重要であることを指摘している [宮崎 99b]。これは、2 章で述べた不完全知覚の問題に関連しており、強化学習において問題視される重要な事項の一つである。提案法においては、環境の状態が滑らかな変化を伴う場合*6であれば、エージェントが不知覚の状態（エピソード途中で逃亡者が見える状態から見えない状態になった場合）に陥っても、AHP の一対比較表でそれまでに蓄積されている情報（これが短期記憶として機能する）を利用する

*6 例えば「右上」にいた逃亡者が次のステップでは突如「左下」にいるような場合ではないことを指す。

ことで各行動の重みを算出するため、適切な行動が選択される傾向をある程度のステップ数において維持することができる。このように、エージェントは不知覚状態においてもより適切な行動を実行し、その状態と行動の対の情報が学習器に送られ学習が進行する(6・2 節参照)。

ポイント(6)に関して:2 章で触れたように、提案法では「基礎知識」か「学習による知識」かのジレンマが発生する。提案法では、AHP 器からの重みと学習器からの重みをある割合により合成する方法(合成比固定法と合成比減衰法)で解決した。7 章の実験結果から観測されたように、基礎知識のみを利用する(つまり、学習により得られる知識を全く利用しない)方法および基礎知識の利用割合を固定する合成比固定法によって最終的に得られる学習結果は合成比減衰法の結果よりも劣っていた(表 8 参照)。この事実から「強化学習において、基礎知識に伴う行動の重みの利用が学習の長期に及ぶと、学習自体に悪影響を与えると共に、最終的に得られる学習の性能を阻害する」と考えられる。よって、合成比減衰法のように AHP 器の重みの利用割合(つまり、基礎知識の利用割合)を学習が進行する過程で徐々に減少させる方法は、基礎知識による学習への悪影響を学習の最終段階では低減できる一方法として有用と考えられる^{*7}。

最後に、提案法の特徴について考える。強化学習は報酬をもたらす行動を優先するように環境への適応を目指す学習制御の枠組みである。このことから、その学習器は大域的かつ長期的な学習機能を有すると解釈できる。一方、提案法において学習器と併用される AHP 器は、状態認識器から与えられる情報にもとづいて、より適切となり得る行動が優先されるように候補となる行動群を重み付けする。これにより、AHP 器は環境における現在または最近の状態にもとづいた局所的かつ短期的な行動の判断メカニズムと共に記憶メカニズムも有すると言える。このようなハイブリッド化によって、学習初期において良いとは言えない学習器の性能を AHP 器によって補うことが可能であり、提案法の大きな特徴の一つである。

9. お わ り に

現実の場面にあらわれるような問題の多くは複雑かつ動的な変化を伴う。そのような問題に対して強化学習を適用し優れたエージェントを提供するためには、強化学習の効率化を念頭におき、より利用しやすい強化学習法の開発が必須である。本論文では、基礎知識を導入することを考え、その手段として階層化意志決定法(AHP)を利用する強化学習エージェントモデルを提案し、Profit Sharing により強化学習を行う方法を示した。

強化学習エージェントは、目標状態に到達した際に与

えられる報酬を手掛かりに学習を行うため、報酬が得られなければ学習は進行しない。その観点に立つ時、例えば複雑なタスクに対しても確実に報酬を得るための有効な方法の一つは、強化学習に対して(人間が予め設定する目標状態に対しての)知識を導入することと考えられる。目標の存在は人間が設定する限り既知であるから、エージェントに導入される知識はその既知である目標に対する基礎的なものと捉えることができ、本論文ではこれを「基礎知識」と定義した。提案法は、マルチエージェントが目標を達成するために各エージェント自身が本来備えていても不思議ではない、人間にとっては明らかである基礎知識を AHP により設計し、従来の学習器と併用することで強化学習の効率化を目指すものである。

その併用の方法として、学習器および AHP 器から得られる行動群の重みの利用率を常に固定する合成比固定法と、学習が進行するにつれて利用率を AHP 器から学習器へ徐々に高める合成比減衰法を示した。追跡問題を対象とした実験の結果、後者の方法の有用性が確認でき、提案法は、従来法に比べて、学習初期の段階から非常に効率の良い学習を実現できることを示した。さらに、合成比減衰法を有する提案法は学習後期においても従来法と同等の学習性能を発揮できることを明らかにした。

強化学習は報酬をもたらす行動を優先するように環境への適応を目指す学習制御の枠組みであることから、その学習器は大域的かつ長期的な学習機能を有すると解釈できる。一方、我々のアプローチで導入される AHP 器は、状態認識器から与えられる情報にもとづき、より適切となり得る行動が優先されるように候補となる行動群を重み付けする機能を持っているので、環境における現在または最近の状態にもとづいた局所的かつ短期的な行動の判断メカニズムと共に記憶メカニズムも有すると言える。これら両器の機能を総合的に考慮すると、提案法は、学習初期において良いとは言えない学習器の性能を AHP 器の機能によって補うことができ、効率的な学習を実現する一種のハイブリッドによる学習手法と見なすことができる。

このような観点から、合成比減衰法を有する我々の方法は、学習の進行に伴い基礎知識の利用割合を徐々に減少させるので、これを汎用的に捉えると、結局、局所的な判断かつ短期的な記憶を重視する学習から大域的な学習機能を重視する方向へ徐々にシフトさせる方法と解釈できる。このようなハイブリッド化により、学習初期では効率的な学習を行い、基礎知識による学習への悪影響を抑制しつつ、学習後期でも良好な学習性能を得ることが可能となった。従って、大域的および局所的な機能を有する我々の強化学習アプローチは、強化学習の効率化・改善に向けた有効な一アプローチであると考えられる。さらに、現実の場面にあらわれるような迅速な対応や安全性・確実性が求められる応用においても汎用的に対応でき得る一手法であると期待できる。

^{*7} 仮に悪影響が低減できないような状況下になったとしても、提案法の学習器では忘却型の強化学習法[加藤 01]などのより高性能なアルゴリズムを利用することも可能である。

今後の課題として、提案法をより複雑な問題へ適用し、適用範囲の可能性に関して明らかにすること、AHP 器と学習器の重みの合成比を学習の状況に応じて適切に決定可能とする適応型の合成法や基礎知識の導入により強化学習に与える影響の詳細について検討する必要がある。

謝 辞

合同エージェントワークショップ&シンポジウム 2003 において、本研究に対しまして有益なコメント等を頂きました、大沢英一先生、荒井幸代先生に感謝致します。

◇ 参 考 文 献 ◇

- [荒井 98] 荒井 幸代, 宮崎 和光, 小林 重信: マルチエージェント強化学習の方法論—Q-learning と Profit Sharing による接近—, 人工知能学会誌, Vol. 13, No. 4, pp. 609–618 (1998)
- [Benda 85] Benda, M., Jagannathan, V., and Dodhiawalla, R.: On optimal cooperation of knowledge sources, *Technical Report BCS-G2010-28, Boeing AI Center, Boeing Computer Services, Bellevue, WA* (1985)
- [Dixon 00] Dixon, K. R., Malak, R. J., and Khosla, P. K.: Incorporating Prior Knowledge and Previously Learned Information into Reinforcement Learning, *Technical Report, Institute for Complex Engineered Systems, Carnegie Mellon University* (2000)
- [Grefenstette 88] Grefenstette, J. J.: Credit assignment in rule discovery systems based on genetic algorithms, *Machine Learning*, Vol. 3, pp. 225–245 (1988)
- [伊藤 01] 伊藤 昭, 金淵 満: 知覚情報の粗視化によるマルチエージェント強化学習の高速化—ハンタゲームを例に—, 信学論 (D-I), Vol. J84-D-I, No. 3, pp. 285–293 (2001)
- [Kaelbling 96] Kaelbling, L. P., Littman, M. L., and Moore, A. W.: Reinforcement learning: a survey, *Journal of Artificial Intelligence Research*, Vol. 4, pp. 237–285 (1996)
- [加藤 01] 加藤 新吾, 松尾 啓志: 動的環境下における Profit Sharing, 信学論 (D-I), Vol. J84-D-I, No. 7, pp. 1067–1075 (2001)
- [松浦 97] 松浦 賢一, 嘉数 侑昇: 非均質エージェント系における組織的行動の形成: 2 次元追跡問題における考察, 情報処理学会論文誌, Vol. 38, No. 6, pp. 1083–1093 (1997)
- [参沢 00] 参沢 匡将, 木村 春彦, 広瀬 貞樹, 大里 延康: 強化学習型マルチエージェントによる交通信号制御, 信学論 (D-I), Vol. J83-D-I, No. 5, pp. 478–486 (2000)
- [宮崎 94] 宮崎 和光, 山村 雅幸, 小林 重信: 強化学習における報酬割当ての理論的考察, 人工知能学会誌, Vol. 9, No. 4, pp. 580–587 (1994)
- [宮崎 99a] 宮崎 和光, 荒井 幸代, 小林 重信: Profit Sharing を用いたマルチエージェント強化学習における報酬配分の理論的考察, 人工知能学会誌, Vol. 14, No. 6, pp. 1156–1164 (1999)
- [宮崎 99b] 宮崎 和光, 木村 元, 小林 重信: Profit Sharing に基づく強化学習の理論と応用, 人工知能学会誌, Vol. 14, No. 5, pp. 800–807 (1999)
- [Ono 96] Ono, N. and Fukumoto, K.: Multi-agent Reinforcement Learning: A Modular Approach, in *Proc. of the Second International Conference on Multi-Agent Systems*, pp. 252–258 (1996)
- [大沢 95] 大沢 英一: 協調プランニングにおける動的組織再編とメタレベル整合戦略—追跡ゲームにおける考察, コンピュータソフトウェア, Vol. 12, No. 1, pp. 43–51 (1995)
- [Russell 97] Russell, S. and Norvig, P.: エージェントアプローチ人工知能, (古川 康一監訳) 共立出版 (1997)
- [Saaty 80] Saaty, T.: *The analytic hierarchy process*, The McGraw-Hill Companies (1980)
- [Sutton 98] Sutton, R. S. and Barto, A. G.: *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA (1998), (邦訳: 強化学習, 三上 貞芳, 皆川 雅章 共訳, 森北

出版, (2000))

- [高橋 01] 高橋 友一, 伊藤 暢浩: RoboCup ではじめるエージェントプログラミング, 共立出版 (2001)
- [刀根 86] 刀根 薫: ゲーム感覚意志決定法—AHP 入門, 日科技連出版社 (1986)
- [Unemi 00] Unemi, T.: Scaling up Reinforcement Learning with Human Knowledge as an Intrinsic Behavior, in *Proc. of the Sixth International Conference on Intelligent Autonomous Systems*, pp. 511–518 (2000)
- [Watkins 92] Watkins, C. J. C. H. and Dayan, P.: Technical note: Q-learning, *Machine Learning*, Vol. 8, pp. 279–292 (1992)
- [山村 95] 山村 雅幸, 宮崎 和光, 小林 重信: エージェントの学習, 人工知能学会誌, Vol. 10, No. 5, pp. 683–689 (1995)

〔担当委員: 北村泰彦〕

2003 年 10 月 30 日 受理

—— 著 者 紹 介 ——

片山 謙吾(正会員)

1992 年 岡山理科大学工学部電子工学科卒業。1995 年 同大大学院修士課程修了。1998 年 同大大学院博士課程修了。博士(工学)。同年同大工学部情報工学科助手。2001 年 同大工学部情報工学科講師。メタヒューリスティクス、組合せ最適化、強化学習に関する研究に従事。電子情報通信学会、情報処理学会各会員。

奥石 尚宏

2003 年 岡山理科大学工学部情報工学科卒業。現在、同大大学院修士課程在学中。強化学習に関する研究に従事。電子情報通信学会学生会員。

成久 洋之

1970 年 京都大学大学院博士課程修了。工学博士。現在、岡山理科大学工学部情報工学科教授。1983 年～1994 年 岡山理科大学情報処理センター所長。オペレーションズリサーチおよびシステムの最適化に関する研究に従事。日本オペレーションズリサーチ学会フェロー。電子情報通信学会、情報処理学会各会員。