複数行動結果を考慮した最尤推定に基づく 状態一般化法

State Generalization Based on Maximum Likelihood Estimation Considering Multiple Behavior Outcomes

矢入 健久

Takehisa Yairi

東京大学先端科学技術研究センター

Research Center for Advanced Science and Technology, Univ. of Tokyo
yairi@ai.rcast.u-tokyo.ac.jp

堀 浩一

Koichi Hori

東京大学大学院工学系研究科先端学際工学専攻

Interdisciplinary Course on Advanced Science and Technology, Univ. of Tokyo
hori@ai.rcast.u-tokyo.ac.jp

中須賀 真一

Shinichi Nakasuka

東京大学大学院工学系研究科航空宇宙工学専攻

Dept. of Aeronautics and Astronautics, Univ. of Tokyo
nakasuka@space.t.u-tokyo.ac.jp

keywords: robot learning, behavior acquisition, state generalization, information entropy

Summary

State generalization problem is a significant issue for the realization of the autonomous agents which are expected to decide and learn the proper behavior with various kinds of sensor information. This paper proposes a new state generalization method based on maximum likelihood estimation of the agent's behavior outcomes. This provides a general framework for unifying the various conventional heuristic generalization criteria which have been used in the previous works, and a way of adapting the state space gradually to the environment.

1. ま え が き

近年、ビヘービアベースドアーキテクチャや、シチュエーテッドエージェントなどに代表されるような適応的、反射的エージェントアプローチが自律知能システム研究の主要分野になりつつある。この研究分野における中心的課題として、“行動獲得問題”，すなわち、自律エージェントが環境との相互作用を通していかに適切な行動規則を学んでいくかという問題があり、実際に様々な機械学習手法を応用した研究が数多く報告されている [Franklin 96, Agre 95]。

しかし、これらの従来研究は多くの場合、エージェントの内部状態集合（状態空間）と行為集合（行為空間）との間のマッピングの獲得という面に主眼を置いている。すなわち、状態空間や行為空間自体は低レベルなセンサー入力・モーター出力空間を離散化・抽象化することによってあらかじめ規定されてしまっている（図1）。このような状態と行為間のマッピング獲得のみに焦点を当てたアプローチには、以下のような重大な限界が存在すると考えられる。

まず第一に、エージェントにとって適切な状態空間や

行為空間を人間があらかじめ定義することは、必ずしも容易でないという点である。すなわち、タスクが高度になり、同時にセンサー入力やモーター出力が増えることによって、環境との関わり合い方が複雑になった場合、その高次のセンサー入力空間やモーター出力空間を人間が直観的に把握し、適切に離散化することは非常に困難になると考えられる。

第二に、仮に適切な状態空間、行為空間を与えられたとしても、環境やエージェント自身のセンサー・アクチュエータの特性の変化によって、それらが無効になってしまう可能性があるという点である。また、事前に定義した状態空間・行為空間がエージェントの全行動獲得過程を通して最適であるという保証もない。

これらの理由により、いかにしてエージェント自身が自律的に状態や行為の一般化を行うか、という状態・行為の自律的一般化問題の重要性が近年認識されるようになり、実際に状態の自律的一般化に関しては既にいくつかの手法が提案されている。

しかし、この状態一般化問題に関しては、依然未解決の問題が多く存在する上に、従来提案されている諸手法

図 1 自律エージェントにおける行動決定と学習過程

間の体系的な関係がほとんど明らかになっていない。これについて、[矢入 99] は、従来の諸手法が「何をもって異なるセンサー入力間の近い・遠いを定めるか」という類似性基準あるいは状態一般化基準、および「同じ状態クラスとして分類されたセンサー入力の集合をどのように一般化し表現するか」という状態一般化・表現手法という二つの観点から分類できることを指摘している。

本稿では、この二つの観点のうち前者の状態一般化基準に焦点を当て、従来の諸方法を包含する新たな一般的枠組として「複数行動結果を考慮した最尤推定に基づく状態一般化法」を提案する。そして実際に数種類の異なる一般化基準を用いてエージェントの状態空間を構成・再構成した場合の行動学習結果について実験的な考察を行う。

以下本稿は次のように構成される。まず2章では従来提案されてきた状態一般化法を概観し、行動結果の類似性に基づく状態の一般化という観点から分類を行う。次に3章では、それら従来手法の包括的な枠組としての「複数行動結果を考慮した最尤推定に基づく状態一般化」と、これに基づく「判別木を用いた状態空間構成法」を提案する。4章では提案手法を自律移動ロボットの行動学習問題に適用した場合のシミュレーション結果を示し、その妥当性を検証する。さらに5章では提案手法における状態数（粒度）の自動決定問題についても考察を行う。

2. 状態一般化問題

2.1 行動結果の類似性に基づく状態の一般化

自律エージェントにおける状態一般化とは、エージェントが観測する様々なセンサー入力パターンのうち類似のものを同一の状態クラスとして一般化することによって、一般に高次元で連続的であるプリミティブなセンサー入力空間を有限個の部分空間に離散化する過程として考えることができる（図2）。

ここで注意すべき点は、このセンサー入力の一般化が基本的には教師無し学習過程であるということである。すなわち、エージェントが経験する各センサー入力パター

図 2 センサー入力空間の離散化と状態の一般化

ンは、それがどの状態クラスに属するかというラベル付けがなされていないので、エージェントは自らセンサー入力パターンを分類するための何らかの基準を持つ必要がある。その最も単純な方法としては、センサー入力空間を均等に、例えば格子状に分割することが考えられる。しかし、例えば、二つのセンサー入力ベクトルが数値上は近くても、同じアクションに対して全く異なる結果（例えば報酬）をもたらす状況が有り得る。この場合にそのような単純なセンサー空間分割を行って状態・行為マッピング獲得を行っても、良い学習結果は得られない。このことが意味するのは、センサー入力パターン間の類似度は、センサー入力値そのものだけでなく、その状態である行為を実行した場合にどのような結果が得られるかということによって決まるということである。

一方、これとは対極的なアプローチとして、学習後のエージェントの行動パフォーマンスの善し悪しを状態空間の評価関数として見なし、これを最適化するようなセンサー空間分割を直接的に探索することが考えられる。しかし、この方法は学習結果の最適性という観点からは確かに理想的ではあるが、膨大な試行回数、すなわち学習のコストを要するという点で非現実的である。

そのため、これまでに提案されてきた状態一般化法の多くは、暗黙的に一つの共通したヒューリスティクスによってセンサー入力パターン同士の距離尺度を定義し、それに基づいて経験によって得られたセンサー入力を分類し、状態クラスとして一般化するというアプローチを取っている。そのヒューリスティクスとは、「同じ（似た）行為を適用したときに、同じ（似た）結果を得るようなセンサー入力同士を近いとみなす」というものである（図3）。この考えに基づいた状態一般化法を、本稿では「行為と行動結果の類似性に基づく状態の一般化」と呼ぶものとする。

2.2 状態一般化基準の違いによる従来手法の分類

一概に「行動結果の類似性に基づく状態一般化」と言っても、これに属する既存の個々の状態一般化手法には、「何をもって行動結果の近い・遠いを定義するか」という点において、いくつかの異なる立場が存在する。この立場の違いは、「状態一般化ポリシー」の違いとも呼べるが、ここではそれらを（1）ゴール・サブゴール到達に基づ

図 3 行為と行動結果の類似性に基づく状態の一般化

く状態一般化 (2) 獲得報酬の類似性に基づく状態一般化 (3) センサー入力変化の類似性に基づく状態一般化の 3 種類に分類し、それぞれの性格について考察・比較する。

§1 ゴール・サブゴール到達に基づく状態一般化

ゴール・サブゴール到達に基づく状態一般化 [Asada 96, Albus 95, Yairi 96] では、まず (あるタスクにおける) ゴール状態のみがエージェントに与えられており、同じアクションによってこのゴール状態に到達できた前状態におけるセンサー入力の集合がサブゴール状態として一般化される。そして、以後これと同様に、同じアクションによって同じサブゴールに到達するセンサー入力集合が新たなサブゴール (状態クラス) として一般化されることでセンサー入力空間が徐々に離散化されていく。

この状態一般化ポリシーの最も重要な性格は、一般化された状態クラスの集合とともに、任意の状態からゴール状態に到達するための行動経路も同時に獲得されるということである。すなわち、このポリシーによって最終的に得られる状態空間は、そのタスクにとっては“最適”であると言える。しかしその反面、二つの大きな問題点も存在する。第一にエージェントがゴール状態あるいはサブゴール状態に到達する行動を経験しない限り新たなサブゴールが生成されないため、学習効率が一般に悪いという点である。これはゴール状態への到達が容易でないタスクにおいては特に重大である。第二に、この一般化ポリシーによって得られた状態クラスの集合は、そのタスクにおけるゴール状態に特化して作られるため、ゴール状態が異なる他のタスクへの再利用が難しいという点である。

§2 獲得報酬の類似性に基づく状態一般化

ゴール状態への到達というイベントは、正の報酬獲得の特殊なケースとして捉えることができる。そこで、上の「ゴール・サブゴール到達に基づく一般化」よりも、より一般的な枠組として、「獲得報酬の類似性に基づく状態一般化」というものが考えられる [Ishiguro 96, 上野 00, Chapman 91]。この一般化ポリシーでは、同じアクションによって環境から類似の報酬を受けるようなセン

サー入力集合を一つの状態クラスとして一般化する。したがって、ゴール・サブゴール到達に基づく状態一般化と比べて、正負の複数の報酬源を考慮することによってより広範なタスクに適用することが可能になる。ただし、報酬源が複数存在する場合、状態の一般化とともにローカルな報酬を獲得するための (あるいはしないための) 行動政策も得られるが、任意の状態からゴール状態に到達するための大局的な行動政策は別途学習する必要がある。

報酬に基づく状態一般化ポリシーは、ゴール・サブゴール到達に基づく状態一般化と同様、学習の初期段階においては直接報酬を獲得する行動経験以外は無駄になるという点で効率が悪く、また、得られる状態クラス集合は報酬の定義に強く依存するため、他のタスクでの再利用が難しいという問題を持つ。

§3 センサー入力変化の類似性に基づく状態一般化

上の二つの状態一般化ポリシーがいずれも“ゴールへの到達”や、“報酬の獲得”という特定のタスクに強く依存した行動結果に基づいているのに対して、これらに依存しない「センサー入力の変化の類似性に基づいた状態の一般化」に属するアプローチも提案されている [Takahashi 96]。この一般化ポリシーでは、同じアクションによって似たセンサー入力変化をもたらす前状態のセンサー入力集合が一つの状態クラスとして一般化される。

この方法の主な長所としては、以下の 2 点が挙げられる。まず第一に、ゴール状態への到達や報酬の獲得などといった特別なイベントと直接関わらない行動経験も、状態の一般化に利用されるため、データ使用の効率が前述の二つの一般化ポリシーと比べて非常に高いという点である。第二に、獲得される状態クラスの集合が特定のタスク、すなわちゴール状態や報酬の定義に依存しないため、他のタスクへの再利用が比較的容易であると考えられる。言い替えば、この一般化ポリシーを用いた場合、極めて“汎用な”状態空間が得られる。しかし、その反面、この汎用な状態集合は、特定のタスクにおけるゴール状態到達や、正報酬獲得行動を学ぶという観点からは、必ずしも最適なものにはならない。すなわち、得られる状態空間は報酬やゴール状態の分布状況を反映するわけではないので、エージェントにとってはあまり興味のない状態クラスも無駄に生成されてしまう可能性がある。

表 1 は、上で考察した 3 つの状態一般化ポリシーの主な特徴を比較して記したものである。これから、いずれの一般化ポリシーもそれぞれ長所・短所を持ち、どれが優れているかという議論よりも、これらをいかにして統合するかを考える方がより重要であると結論づけられる。特に、行動政策獲得、つまり状態・行為マッピング学習との連携を考えた場合、まだ適切な行動政策がほとんど獲得されていない学習の初期段階においては経験データ使用率の高いセンサー入力変化の類似性に基づいたポリシーに基づいて状態一般化が行われ、逆にある程度行動

表 1 3つの状態一般化ポリシーの比較

	(サブ)ゴール到達 に基づく一般化	報酬獲得の類似性 に基づく一般化	センサー変化の類似 に基づく状態一般化
状態生成の様子	ゴール状態から徐々に	報酬から徐々に	全領域同時並行
行動政策の入手性	同時に得られる	部分的に得られる	得られない
タスク依存性	非常に高い	高い	低い
データ利用率	非常に低い	低い	高い
結果の再利用性	非常に低い	低い	高い
状態空間の最適性	高い	中程度	低い
全体的性格	ゴール指向	報酬指向	データ駆動型

政策が学習されてからはそのタスク固有のゴール状態あるいは報酬に基づいて状態クラスが再定義されていくことが理想的であると考えられる。

この背景に基づき、次章ではこれらの異なる状態一般化ポリシーを統合する新たな枠組を提案する。

3. 複数行動結果を考慮した最尤推定に基づく状態の一般化

前章では、既存の状態一般化法の多くが、同じアクションを行ったときの行動結果の同一性あるいは類似性に基づいてセンサー入力の集合を分類し一般化するというアプローチを取っていること、そしてその過程において何を行動結果として扱うかによって一般化される状態空間の性格に違いが生じることを述べた。本章では、これらの異なる状態一般化ポリシーを統合する新たな枠組として、複数行動結果を考慮した最尤推定に基づく状態の一般化を提案する。

3.1 基本的アイデア

前章で述べたように、従来の行動結果の類似性に基づく状態一般化法に共通する考え方は、「同じアクションを行ったときに同じ（あるいは類似の）行動結果が得られるようなセンサー入力を一般化する」というものであった。これは「あるアクションが適用されたときの行動結果をできるだけ正確に予測することができるようにセンサー入力を一般化し、離散的な状態集合を得る」過程として解釈することができる。

そこで本研究では、ある状態集合を用いてエージェントの行動結果予測モデルを過去の行動経験に基づき最尤推定した場合の尤度の大きさをその状態集合の評価関数として定義し、これを最大化するような状態集合、すなわちセンサー入力空間の分割を求める問題として「行動結果の類似性に基づく状態一般化」を定式化する。

3.2 最尤推定に基づく状態一般化問題の定式化

まず、エージェントの行動経験 beh が、行為前のセンサー入力ベクトル s 、実行された行為 a 、そしてこの行為によって得られた行動結果ベクトル r の3つの要素の組によって表現されるものとする ($beh = \langle s, a, r \rangle$)。セ

ンサー入力ベクトル s はエージェントのセンサー入力ベクトル空間 $S (= S_1 \times \dots \times S_{d_s})$ 中の1点で表され、行為 a はあらかじめ定められた行為集合 $\mathcal{A} = \{A_1, A_2, \dots, A_{n_a}\}$ の要素のいずれかを値として取る。また、行動結果ベクトル r も行動結果ベクトル空間 $\mathcal{R} (= R_1 \times \dots \times R_{d_r})$ 中の1点で表されるが、その構成次元としては、行為後のセンサー入力ベクトル、獲得した報酬値、(以前の状態空間における) 到達した状態クラスなど、任意の行動結果属性を含めることができるものとする。

今、エージェントの1つの行動経験 $beh = \langle s, a, r \rangle$ のうち、センサー入力 s と行為 a の値が与えられた時に行動結果 r の値を以下のように予測する行動結果予測モデル $M = \langle \mathcal{C}, \mathcal{A}, \Theta \rangle$ を考える。まず、 M はエージェントのセンサー入力空間 S を離散的な状態クラス集合 $\mathcal{C} = \{C_1, C_2, \dots, C_{n_c}\}$ に分割し、 s をその要素 C_k に分類する。また行為 a も行為集合 $\mathcal{A}$ 中のいずれかの要素 A_j に分類される。そして、 C_k と A_j の組合せに対して行動結果 r の確率分布を与えるパラメータを $\theta_{k,j}$ とし、全パラメータの集合を $\Theta = \{\theta_{1,1}, \dots, \theta_{n_c, n_a}\}$ とする。

このとき、エージェントの過去の有限の行動経験集合 $B = \{beh_1, beh_2, \dots, beh_{n_B}\}$ から行動結果予測モデル M を最尤推定する問題を考える。まず、 B に対するモデル M の対数尤度 $LL(M|B)$ は次のように表される。

$$LL(M|B) = \log P_M(B) = \sum_{k=1}^{n_c} \sum_{j=1}^{n_a} \log P_{\theta_{k,j}}(B_{k,j}) \quad (1)$$

ただし、 $B_{k,j}$ は B に含まれる要素のうち、センサー入力 s が状態クラス C_k に分類され、行為 a が A_j であるような行動経験の集合である ($B_{k,j} = \{beh | s \in C_k, a = A_j, beh \in B\}$)。また、 P_M はモデル M が与える確率分布であり、 $P_{\theta_{k,j}}$ はこのうち C_k と A_j の組合せに付随するパラメータ $\theta_{k,j}$ が与える確率分布である。

ここで、各 $B_{k,j}$ 中の行動経験データが互いに独立であると仮定し、 $B_{k,j}$ に含まれるデータ数を $n_{B_{k,j}}$ で表すと、次の式が成り立つ。

$$\log P_{\theta_{k,j}}(B_{k,j}) = \sum_{i=1}^{n_{B_{k,j}}} \log P_{\theta_{k,j}}(r_i) \quad (2)$$

以下、簡単のため行動結果ベクトル r が1次元 ($r = [r]$) で、かつ離散値 $V_1, \dots, V_{n_r}$ を取る場合を考える。したがっ

て確率パラメータベクトル $\theta_{k,j}$ は $\theta_{k,j} = [\theta_{k,j,1}, \dots, \theta_{k,j,n_r}]$ となる ($\sum_m \theta_{k,j,m} = 1$)。このとき、式 (2) は次のように変形される。

$$\sum_{i=1}^{n_{B_{k,j}}} \log P_{\theta_{k,j}}(r_i) = n_{B_{k,j}} \cdot \sum_m P_{B_{k,j}}(r = V_m) \cdot \log \theta_{k,j,m} \quad (3)$$

ここで $P_{B_{k,j}}$ は経験サブ集合 $B_{k,j}$ 内での行動結果 r に関する経験確率分布 (頻度確率分布) である。式 (3) は各 $m (= 1, \dots, n_r)$ について $\theta_{k,j,m} = P_{B_{k,j}}(r = V_m)$ としたときに最大値を取り、その値は $-n_{B_{k,j}} \cdot H_{B_{k,j}}(r)$ になる。ただし、 $H_{B_{k,j}}(r) = -\sum_m P_{B_{k,j}}(r = V_m) \cdot \log P_{B_{k,j}}(r = V_m)$ は、行動経験部分集合 $B_{k,j}$ 内における行動経験 r に関するエントロピーである。したがって、これと式 (1)、(2) より、ある状態クラス集合 C を仮定したときのモデルの最大対数尤度は、 $LL^*(M|B) = -n_B \cdot H_{C,A}(r)$ になる。ただし、 $H_{C,A}(r)$ は全行動経験集合 B を状態クラス集合 C および行為集合 A によって分類した場合の行動結果 r に関する平均エントロピーである。ここで一般に r が多次元ベクトルである場合 ($r = [r_1, \dots, r_{d_r}]$) も、その要素が互いに独立であると仮定すれば、

$$LL^*(M|B) = -n_B \cdot \sum_{l=1}^{d_r} H_{C,A}(r_l) \quad (4)$$

となる。

この最大対数尤度 $LL^*(M|B)$ を状態クラス集合 C に関する評価関数 $J_{B,A}(C)$ として定義することによって、「最尤推定に基づく状態一般化」は次のように定式化される。

[定義] (最尤推定に基づく状態一般化)

エージェントの行為集合 A 、行動経験集合 B 、状態クラスの数の最大個数 $n_{c,max}$ がそれぞれ与えられたときに、上の評価関数 $J_{B,A}(C) = LL^*(M|B)$ を最大化するような状態クラス集合、すなわちセンサー入力空間 S の分割の集合 C_{opt} を求める。

式 (4) から明らかなように、これは r の各行動結果変数に関する平均エントロピーの和を最小化する状態クラス集合を求めることと等価である。

3.3 判別木を用いたセンサー入力空間分割

前節において一種の最適化問題として定式化された状態一般化の枠組を実際に適用するためには、主に以下の 2 つの問題を解決する必要がある。一つは、状態クラス集合 C を実際にどのように表現するかという問題であり、もう一つは前述の評価関数 (最大対数尤度) を最大化する C をどのようにして探索するかという問題である。

これらは言い替えば帰納バイアスの選択という一般的な問題であり、様々な選択肢が存在するが、本研究で

図 4 状態判別木

は判別木 (classification tree) を利用して $J_{B,A}(C)$ を貪欲的に局所最大化しながらセンサー入力空間を分割し、状態クラスを定義していくアルゴリズムを提案する。この手法では、図 4 のように、センサー入力の値に関して記述された判別木によって状態クラス集合が表現され、 $J_{B,A}(C)$ の増加値の大きさに基づいて木中のノードが展開されることによって、センサー入力空間が徐々に分割されていく。

[判別木を用いた対数尤度の局所最大化に基づく状態一般化アルゴリズム]

入力: $A, B, n_{c,max}$ 出力: C

- (1) 全センサー入力が入力空間の一つの状態クラスに属するものとして初期化する ($C = \{C_1\}$, $n_c = 1$)。状態判別木はただ一つのリーフから成る。
- (2) 行動経験集合 B 中の全要素をアクション集合 A と現在の状態クラス集合 C に基づいて B のサブ集合の集合 $\{B_{k,j}\}_{k=1, \dots, n_c, j=1, \dots, n_a}$ に分類する。そして各 $B_{k,j}$ について $J_{B_{k,j}} = -n_{B_{k,j}} \cdot H_{B_{k,j}}(r)$ を計算する。
- (3) 各状態クラスに対応する状態判別木中のリーフごとに $J_{B,A}(C)$ ($J_{B_{k,j}}$) の値を最も増加させる分割 (分割されるセンサー入力と分割値との組) を探す。そしてその増加値が最も大きい分割を採用して判別木を展開する。(すなわち新しい状態クラスを定義する)
- (4) 状態クラスの数が $n_{c,max}$ に達したら終了。そうでなければ 2 に戻る。

この状態一般化アルゴリズムは、最尤推定法の局所解探索の欲張りアルゴリズムの一種として見る事ができる一方、広く知られる決定木を用いた教師付き学習アルゴリズム (例えば ID3[Quinlan 84] や C4.5[Quinlan 95]) においてスカラー量のクラス変数に関するエントロピーの局所最小化を行動結果ベクトルのエントロピー最小化を扱うように拡張したものと解釈することもできる。このような判別木・決定木を用いた局所解探索法の一般的な性質として、解空間を狭めることで探索コストを大きく減らすことができる反面、局所解に陥りやすいという短所がある。

また、上で提案したアルゴリズムでは、最終的な状態

図5 状態空間構成と行動政策獲得との統合

クラス数 $n_{c,max}$ はあらかじめ与えられるものとしているが、これをどのようにして決めるべきかという粒度決定の問題が存在する。加えて、判別木を用いたアルゴリズム以外にも、同じ最大対数尤度を評価関数として用いつつ、ボトムアップにセンサー入力をクラスタリングしていくことによって状態一般化を行うアプローチなども考えられる。これらの問題については、5章においてさらに考察する。

3.4 行動政策獲得アルゴリズムとの結合

上で提案した状態一般化アルゴリズムは、行動経験データ集合 B を入力とし、一般化された状態クラス集合 C を出力とする一つの独立したモジュールとして扱うことができる。したがって、Q-learningなどの通常の行動政策獲得法と結合し、両者を交互に適用することが考えられる(図5左)。

ここで重要なのは、この結合によって状態空間の性格が、前述したような「データ駆動型」から「ゴール・報酬指向型」へと自然に遷移していくと期待される点である。例えば、適切な行動則がまだ十分に得られていない初期の段階では、正の報酬を獲得するような行動経験が乏しいため、「センサー入力変化の類似性」や「負の報酬の獲得」に基づいて状態の一般化が行われる。それに対して、適切な行動則が次第に獲得されるに従って、「正報酬の獲得」や「サブゴール到達」に基づいた状態の一般化が行われるようになって考えられる。

ただし、状態空間を構成し直す際には、再構成前の状態空間を用いて学習したQ値($Q^{old}(C_i^{old}, A_j)$)を再構成後の状態空間のためのQ値($Q^{new}(C_i^{new}, A_j)$)に変換する必要がある。これについては、本研究では[Takahashi 96]で提案されている方法を用い、以下の式に基づいて変換を行うものとする。

$$Q^{new}(C_i^{new}, A_j) = \sum_{i'}^{n_{c,old}} freq(C_{i'}^{old} | C_i^{new}) \cdot Q^{old}(C_{i'}^{old}, A_j) \quad (5)$$

図6 想定するロボットと利用可能なアクション

図7 想定するセンサー入力とタスク(報酬)

ここで、 $freq(C_{i'}^{old} | C_i^{new})$ は、新状態クラス C_i^{new} に属するセンサー入力が入り旧状態クラス $C_{i'}^{old}$ にも属する頻度確率であり、経験データ B から推定される。

また、提案する状態一般化アルゴリズムは、前述のようなランダムな行動選択による行動経験収集から始める学習戦略だけではなく、人間が暫定的に与えた初期状態空間を用いて行動政策学習を行いつつ、そのときの行動経験に基づいて状態空間を繰り返し再構成するという学習戦略(図5右)にも適している。このような初期状態空間の利用によって、エージェントの行動経験収集に伴う学習コストを大きく減らすことが期待される。

これらの事項については、次章のシミュレーション結果で具体的に説明する。

4. シミュレーション

4.1 想定するエージェント、タスクおよび環境

本章では、前章で提案した状態一般化法の有効性を検証するために実施した計算機上でのシミュレーションによる実験の結果を説明する。

ここで想定するロボットエージェントは、左右2つの車輪の回転数を指定することによって図6に示すような、8つの行為($A_1 - A_8$)を選択実行することができるものとする。また、センサー入力に関しては、2種類の情報、すなわち、ロボットの重心位置から目標物までの距離(S_1)、ロボットの機軸に対する目標物の相対方向角(S_2)を利用できるものとする(図7)。このロボットが学習すべきタスクとしては、目標物到達行動を想定し、ランダムに与

えられた初期地点を出発して機体の前面が目標物に到達することによって正の報酬 (+10) を与えられるか、決められた領域の外に出てしまうことによって負の報酬 (-10) を与えられるかするまでを 1 つのエピソードとする。

次節で述べるように、本シミュレーションでは、何種類かの行動結果ベクトル r を仮定した場合に対して提案手法に基づき状態空間を構成・再構成した結果を比較するが、いずれの場合も行動政策 (状態 - 行為マッピング) の獲得に関しては、通常の Q-learning を用いるものとする。すなわち、Q 値は次式によって更新される: $Q_t(c, a) = (1 - \alpha)Q_{t-1}(c, a) + \alpha(rwd + \gamma \max_{a' \in A} Q_{t-1}(c', a'))$ 。そして、各時点での行為は現在の Q 値を用い Boltzmann 分布に基づいて確率的に決定される。すなわち、各行為が選択される確率を $P(a|c) = \frac{\exp(Q_t(c, a)/T)}{\sum_{a' \in A} \exp(Q_t(c, a')/T)}$ によって計算し、ルーレット戦略によって取るべき行為を決定する。ただし、各パラメータの値としては以下のものを用いる。 $\alpha = 0.05$, $\gamma = 0.5$, $T = 0.01$ 。

次々節以降の各シミュレーション結果では、学習後のエージェントの行動パフォーマンスの善し悪しを「スタート地点から 30 ステップ以内に目標地点に到達できた割合」の大きさによって定義する。また、グラフに示される値はいずれも 100 回の試行結果を平均したものである。

4.2 想定する状態空間の構成・再構成パターンと状態一般化基準

§1 想定する構成・再構成パターン

今回のシミュレーションでは、移動ロボットエージェントの行動政策獲得 (Q 学習) プロセスと、提案手法に基づく状態空間構成・再構成プロセスの組合せのパターンとして以下の 5 つを想定した。

- P1: ランダムアクションによって行動経験を収集した後、提案手法に基づき状態空間を自律構成。その後 Q 学習で行動政策獲得。
- P2: 上の P1 の後、さらに状態空間を再構成し再度 Q 学習を行う。
- P3: 上の P2 の後、さらに状態空間を再構成 (計 3 回の状態空間構成・再構成)。
- P4: 人間が事前に与えた初期状態空間を用いて Q 学習を行った後、状態空間を再構成する。
- P5: P4 で用いた初期状態空間を使い続ける (再構成なし)。

なお、いずれの場合も状態空間構成・再構成で定義される状態クラスの数 ($n_{c, max}$) は 25 とした。

§2 比較する状態一般化基準

「エージェントが何に基づいてセンサー入力間の距離尺度を定め、状態を一般化するか」という状態一般化基準は、2 章、3 章で述べたように、本研究の枠組においては「何を行動結果ベクトルの要素として選択するか」ということに帰着される。

図 8 ランダムアクションからの状態空間構成時 (P1-C1, P1-C2) におけるパフォーマンス変化

そこで、今回のシミュレーションではこの行動結果ベクトル r の要素として、まず以下のものを考える。

- 行為後のセンサー入力 (s_{post})
- 直接獲得報酬 (rwd)

また、既存の状態空間を再構成する場合には、これに加えて以下の 2 つの行動結果要素も用いるものとする。

- 状態遷移に伴う間接報酬も含んだ獲得報酬 ($rwd' = rwd + \gamma \max_{a' \in A} Q_{t-1}(c', a')$)
- 行為によって遷移した (再構成前の状態空間における) 状態クラス (c_{post})

そしてこれらの組合せによって、以下の 5 つの状態一般化基準 (C1-C5) を仮定し、比較を行った。

- C1: s_{post} のみ ($r = [s_{post}]$)
- C2: s_{post} および rwd ($r = [s_{post}, rwd]$)
- C3: c_{post} のみ ($r = [c_{post}]$)
- C4: c_{post} および rwd ($r = [c_{post}, rwd]$)
- C5: rwd' のみ ($r = [rwd']$)

ここで、C3, C4, C5 の状態一般化基準は、既存状態空間から再構成を行う場合 (P3, P4) でのみ利用可能である。

4.3 ランダムアクションによる行動経験収集に基づく状態空間の構成

まず、事前にエージェントに初期状態空間が与えられず、ランダムアクションによって行動経験を収集した後、これに基づいて状態空間を自律構成・再構成する場合 (P1, P2, P3) について実験を行った。

図 8 は、状態空間構成・再構成パターン P1 において、行為後のセンサー入力の値 (s_{post}) のみを考慮した場合 (P1-C1) と、行為によって得られる直接報酬 (rwd) も考慮に入れた場合 (P1-C2) のエージェントのパフォーマンス変化を比較して示したグラフである。状態空間をそれぞれの状態一般化基準に基づいて構成した後 (150 エピソード以降) は、P1-C2 の方が P1-C1 よりも成績が良くなっており、一般化基準に直接獲得報酬を含めるか否かの差が出ていると考えられる。図 9 と図 10 はそれぞれのケースにおいて構成された状態空間の典型例

図 9 P1-C1 において構成された状態空間

図 12 P2-C2 において再構成された状態空間

図 10 P1-C2 において構成された状態空間

図 13 グリッド状（初期）状態空間

図 11 ランダムアクションからの状態空間構成・再構成時におけるパフォーマンス変化

を示したものであるが、直接獲得報酬を考慮した後者のほうが、考慮していない前者よりも正負の報酬源に近い領域においてより細かく状態が分割されていることから裏付けられる。なお、最初の 150 エピソードにおいていずれのケースともパフォーマンスが著しく低くなっているのは、行動経験の収集のためにランダムアクションを行っているためである。

次に、状態一般化基準 C2 を用いた場合について、最初の状態空間構成後に再び再構成を行ったケース (P2-C2)、および、さらにもう一回再構成を行ったケース (P3-C2) でのパフォーマンス変化を調べた。図 11 はこれら 2 つのケースと先の P1-C2 とを比較したものであるが、Q 学習時に集められた行動経験データを用いて状態空間を再構成した後 (それぞれ 400 エピソード後と 650 エピソード後) に、パフォーマンスが大きく改善されていることが観察される。これは Q 学習による行動政策獲得によって、ゴールに到達するような行動経験が相対

的に増えるために、よりゴール指向の状態空間が形成されるためであると考えられる。実際、図 12 は P2-C2 における再構成後の状態空間の様子を示したものであるが、ゴール状態 (正の報酬源) に近い領域 (センサー 1 の値が 50 以下でセンサー 2 の値が 0 付近の領域) においてより微細に分割されている。

4.4 初期状態空間を利用した再構成

前節で用いた状態空間構成・再構成パターン P1, P2, P3 では、最初にランダムアクションによって行動経験を収集し、ゼロから状態空間の構成を行った。しかし、このような膨大な試行錯誤を要する学習過程を実環境で実施することはあまり現実的ではない。

そこで、3.4 節で述べた考えに基づき、図 13 に示した格子状の状態空間 (状態数 25) を事前に与えて Q 学習を行った後に状態空間を再構成する戦略 (構成・再構成パターン P4) の有効性を調べた。図 14 は、この状態空間再構成戦略を状態一般化基準 C2 とともに用いた場合 (P4-C2) と、同じ一般化基準を用いつつゼロから状態空間を構成・再構成した場合 (P2-C2) のパフォーマンス変化を示したグラフである。比較のために、図 13 の格子状状態空間を使用し続けた場合の結果も同時に示してある。この結果より、適切な初期状態空間を与えることが可能な場合には、それを利用して行動政策獲得を行った後に状態空間を再構成することで、より適切な状態空間を獲得するために必要な試行錯誤の回数 (すなわち探索コスト) を大きく減らすことが可能になることがわかる。

次に、同じ学習・状態再構成パターン P4 を採用しつつ、前述の数種類の状態一般化基準 (C2-C5) を用いた場合におけるそれぞれの行動パフォーマンスを比較した

図 14 初期状態空間を利用した再構成時 (P4-C2) におけるパフォーマンス変化

図 15 異なる状態一般化基準におけるパフォーマンス変化

(図 15). この実験より, このタスク環境においては, 状態一般化基準に前状態空間における遷移先状態 (c_{post}) と直接報酬 (rw_d) を含めて再構成を行った場合 (C4) が, c_{post} のみを考慮した場合 (C3) や, 直接・間接報酬 (rw_d') のみを考慮した場合 (C5) よりもエージェントのパフォーマンスが良くなるという結果が得られた.

5. 提案手法の限界と展望

5.1 粒度 (状態数) の自動決定問題

3.3 節で提案した判別木を利用した状態一般化アルゴリズムでは, 最終的な状態 (クラス) の数を事前に与えることになっているが, これは必ずしも自明なことではない. すなわち, 指定した状態数が少なすぎる場合には, 本質的に別々に扱うべき状態を同一のものとして扱ってしまうために適切な行動を学べない可能性が生じる. 一方, 状態数が多すぎる場合, 行動政策 (状態 - 行為マッピング) の獲得に膨大な時間とコストがかかることになる. また, 状態空間を必要以上に精密化することは, 経験データ中のノイズに特化してパフォーマンスを低下させる可能性をもたらすと考えられる.

そこで, 実際に前章の再構成シミュレーション (P4-C4) において, 判別木によって生成する状態クラスの数を増減させた場合に最終的な (1000 エピソード後の) 行動パフォーマンスがどのように変化するかを調べた (図

図 16 状態数を変えた時のパフォーマンス比較

16). この場合, 状態数が約 60 ~ 80 個のときにパフォーマンスが最大になっており, それ以上状態数を増やして評価関数 $J_{B,A}(C)$ の値を大きくしてもパフォーマンス自体は改善されない (むしろ若干低下する) ことがわかる. 他の状態一般化基準 (C1 - C3, C5) を用いた場合も, 同様の傾向が観察された.

このような観点から状態分割数を適切に自動決定する方法として, AIC (赤池情報量基準) や MDL (最小記述長) に代表されるモデルの複雑さとデータに対する尤度とのトレードオフを考慮した情報量基準の利用が挙げられる. 実際, MDL を判別木 (決定木) における枝刈り [Quinlan 89] や分岐停止条件 [鈴木 95] の基準として適用した例は数多く知られており, これらの手法を先に提案した状態分割アルゴリズムに組み入れることは有望なアプローチであると考えられる.

また, 提案手法のように判別木を用いてトップダウンにセンサー空間を分割していくのではなく, あらかじめ細かく離散化された初期状態集合を行動結果 (直接報酬) の類似性に基づいてボトムアップにクラスタリングしていく方法も提案されている [Abe 99]. この方法では, 状態クラスターの粒度決定問題を AIC や MDL を用いて解決しており, これを本研究の問題設定に適用し, 判別木を用いた場合と比較評価することも興味深い課題である.

しかしその一方で, 現時点ではこれらの情報量基準を最適化することとエージェントの行動パフォーマンス (前章のシミュレーションではゴール到達率) や学習速度などの実際の評価基準との間の明確な関係が明らかになっていない. この問題に関しては, 今後, 理論的な解析と実験的な検証が並行して進められていくことが望まれる.

5.2 マルコフ性との関連性

本研究で提案した「複数行動結果を考慮した最尤推定に基づく状態の一般化」という枠組は, 反射的な行動学習エージェントにとって「望ましい」状態空間とは「同じ状態において同じ行為を実行したときに, できるだけ同じ (近い) 結果が期待されるような状態空間である」という一種のヒューリスティクスに基づいている.

しかし、そもそも本研究で用いた Q 学習などの強化学習理論を適用すること自体が妥当か否かという観点からすれば、環境からの状態信号 (state signal) がマルコフ性を満たすかどうかということの方がより根本的な問題であると言える [Sutton 98]。

この問題に関して、本稿では一般化を行う前のセンサー入力空間 S がマルコフ性を満足しており、かつ、いかなる離散の状態空間 C への分割に対してもそのマルコフ性が保存されるという仮定を置いていた。そして、この仮定の上で、「できるだけ状態遷移過程のエントロピーが小さくなるような状態空間 (センサー空間分割) C を求める」という問題を扱っていたと言える。しかし、このマルコフ性の保存に関する仮定は特別な場合を除いて成立しない。すなわち、状態空間 C の定義の仕方によっては非マルコフ性が生じてしまう可能性がある。したがって、今後、マルコフ性を保存しつつ前述の評価関数 (対数尤度) を最大化するような手法の研究が必要となる。また、もともとのセンサー入力空間においてマルコフ性が成立しない場合、最小限の過去の状態への参照 (記憶) を加えるなどして、マルコフ性を改善するような状態空間構成法の拡張も望まれる。

6. む す び

本稿では、自律エージェントの行動学習における状態一般化問題において、特に状態一般化基準の問題に焦点を当て、従来の諸研究において用いられて来た一般化基準を包括する新たな枠組として「複数行動結果を考慮した最尤推定に基づく状態一般化法」を提案した。また、シミュレーションでは、実際に様々な一般化基準を用いて状態一般化を行った場合のエージェントの学習結果の違いや、初期状態空間を利用して Q 学習を行った後に状態空間を再構成することによって得られる効果などを実験的に示した。

今後の課題としては、前章で述べた自動粒度決定、判別木以外の状態表現法への適用、マルコフ性保存の考慮などの他に、提案手法を行為 (アクション) の自律的一般化にも適用できるように拡張するというテーマがある。本稿では、行為集合はあらかじめ人間が定義したものをを用いているが、原理的にはモーター空間の分割にも最尤推定に基づくアプローチが適用可能であると考えている。ただし、実際のロボットなどにおけるアクションは、一般的には、より単位の小さいアクション要素の時系列的な組み合わせによって構成されると考えられ、時間軸も含めたモーター空間の抽象化に拡張する必要がある。

謝 辞

本論文に対して査読者の方々から多数の有益な御助言を頂いた。ここに感謝の意を表したい。なお、本研究の一部は文部省科学研究費補助金 (課題番号 08004297,

12750204) によって実施された。

◇ 参 考 文 献 ◇

- [Abe 99] Abe, N. and Nakamura, A.: Learning to Optimally Schedule Internet Banner Advertisements, in *Proceedings of the Sixteenth International Conference on Machine Learning* (1999).
- [Agre 95] Agre, P. E. and S. J. Rosenchein, eds.: Special Volume: Computational Research on Interaction and Agency, (Part 1, Part2), *Artificial Intelligence*, Vol. 72, 73, (1995).
- [Albus 95] Albus, J., Lacaze, A., and Meystel, A.: Multiresolutional Intelligent Controller for Baby Robot, in *Proceedings of the 10th International Symposium on Intelligent Control* (1995).
- [Asada 96] Asada, M., Noda, S., and Hosoda, K.: Action-Based Sensor Space Categorization for Robot Learning, in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1502–1509 (1996).
- [Chapman 91] Chapman, D. and Kaelbling, L. P.: Input Generalization in Delayed Reinforcement Learning: An Algorithm And Performance Comparisons, in *Proceedings of Twelfth International Joint Conference on Artificial Intelligence*, pp. 726–731 (1991).
- [Franklin 96] Franklin, J. A., Mitchell, T. M., and S. T. Thrun, eds.: Special Issue on Robot Learning, *Machine Learning*, Vol. 23, No. 2/3 (1996).
- [Ishiguro 96] Ishiguro, H., Sato, R., and Ishida, T.: Robot Oriented State Space Construction, in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1496–1501 (1996).
- [Quinlan 84] Quinlan, J. R.: Learning Efficient Classification Procedures and Their Application to Chess End Games, in Michalski, R., Carbonell, J., and Mitchell, T. eds., *Machine Learning: An Artificial Intelligence Approach, Symbolic Computation*, chapter 15, pp. 463–482, Springer-Verlag (1984).
- [Quinlan 89] Quinlan, J. R. and Rivest, R. L.: Inferring Decision Trees Using the Minimum Description Length Principle, *Information and Computation*, Vol. 80, (1989).
- [Quinlan 95] Quinlan, J. R.: AI によるデータ解析, トッパン (1995).
- [Sutton 98] Sutton, R. S. and Barto, A. G.: *Reinforcement Learning*, MIT Press (1998).
- [Takahashi 96] Takahashi, Y., Asada, M., and Hosoda, K.: Reasonable Performance in Less Learning Time by Real Robot Based on Incremental State Space Segmentation, in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1518–1524 (1996).
- [Yairi 96] Yairi, T., Nakasuka, S., and Hori, K.: Automatic Acquisition of Reactive Behavior for Planetary Rovers, *Journal of Space Technology and Science*, Vol. 12, No. 1, pp. 17–26 (1996).
- [上野 00] 上野, 中須賀, 堀: 自律エージェントのための状況認識と行動規則の同時学習, 人工知能学会誌, Vol. 15, No. 2, pp. 297–308 (2000).
- [矢入 99] 矢入, 中須賀, 堀: 異種冗長なセンサー情報に基づく自律的状态抽象化法, 人工知能学会誌, Vol. 14, No. 4, pp. 91–102 (1999).
- [鈴木 95] 鈴木, 圓川: MDL 基準による判別木の生成, 人工知能学会誌, Vol. 10, No. 4, pp. 572–579 (1995).

〔担当委員: 小林重信〕

1999 年 11 月 29 日 受理

著 者 紹 介

矢入 健久(正会員)

1971 年生まれ．1994 年東京大学工学部航空宇宙工学科卒業．1996～1998 年日本学術振興会特別研究員．1999 年東京大学院博士課程修了．同年，東京大学先端科学技術研究センター助手．工学博士．1999 年度本学会論文賞受賞．

堀 浩一(正会員)

1956 年生まれ．1979 年東京大学工学部電子工学科卒業．1984 年同大学博士課程修了．1984 年国立大学共同利用機関国文学研究資料館助手．1986 年同助教授．1988 年東京大学先端科学技術研究センター助教授．1992 年同工学系研究科助教授．1997 年同教授．工学博士．電子情報通信学会，情報処理学会，日本ソフトウェア科学会，日本認知科学会，IEEE，ACM，各会員．

中須賀 真一

1961 年生まれ．1983 年東京大学工学部航空学科卒業．1988 年同大学院博士課程修了．同年，日本アイ・ピー・エム（株）入社．1990 年東京大学工学部航空学科講師．1993 年同助教授．工学博士．AIAA，日本航空宇宙学会，計測自動制御学会，各会員．