餌をくれるのは誰？
～ 不完全知覚問題における困難さの指標の導入 ～

*植村 渉
龍谷大学 理工学部 電子情報学科
〒 520-2151 滋賀県大津市瀬田大江町横谷 9-5
wataru@rins.ryukoku.ac.jp
http://friede.elec.ryukoku.ac.jp/~wataru

Abstract: 人工知能の研究では，生物の挙動からヒントを得て，モデルやアルゴリズムを構築することがある．本研究では，我が家で飼っている猫の学習を基に，学習方法について検討を行う．強化学習の枠組みでは，簡単な環境のモデルとしてマルコフ性のある環境（MDPs）を用い，複雑な環境としては隠れマルコフモデルに基づく環境や，部分観測可能マルコフ決定下（POMDPs）の環境を検討することが多い．MDPs 環境下では，未来の挙動が現在の状態だけで決定される．また，POMDPs 環境では，環境自身は MDPs 環境であるが，学習者であるエージェントの知覚能力に制限があり，本来異なる状態を同一として観測する．これを不完全知覚問題と言う．簡単に言うと，最適な行動選択に必要な入力情報を得られない環境である．例えば，迷路において自分の周り数マスしか観測できない場合，異なる複数の場所が同じように見えることがある．しかし，数歩手前のマスから数えながら移動してみると，それぞれが別のマスであると認識できたりする．このように履歴を用いることで区別する方法が提案されているが，どれだけの履歴の長さを必要とするかは問題環境に依存する．状態遷移が確率的な場合，履歴だけでは一意に状態を決定できなくなる．このときは，区別を諦め，その環境を受け入れる必要がある．代表的な強化学習法である Q-Learning は，MDPs 環境下での学習を想定している．その発展系である Sarsa や報酬分配型の強化学習法である Profit Sharing は，一部の POMDPs 環境下でも学習を進めることができ，実世界への適用が期待されている．しかし，一言で POMDPs 環境といっても，知覚能力に制限のない MDPs 環境と同一の環境から，知覚がなにもできない環境という最も困難なものまで含まれている．本研究では，猫に対する餌付けを通し，生物がどのように不完全知覚問題に対処しているかを検討する．

1. はじめに

強化学習では，学習者であるエージェントを中心とし，問題環境をセンサーなどの入力を通して知覚し，モータやスピーカなどの出力機器を動かすことで問題環境に作用する（図 1）．このとき，エージェントの外部の環境を**状態**，知覚できる観測情報を**観測**と呼ぶ．センサーへのノイズや，モータ出力のすべりなど，エージェント自身が持つ不確実性も，外部環境として定義することができ，正に学習者を中心としたモデルを構築することができる．

環境のモデルとしてマルコフ性を持つ環境（MDPs 環境）を仮定することが多い．そして，状態と観測との写像をどのように扱うかにより，問題の本質が変わってくる（図 2）．

複数の状態が，エージェントにとって本質的に同一であるのであれば，それらをまとめて扱う方が，効率的な学習が期待できる．例えば，サッカーのプレイヤーの行動選択では，フィールドを均等に区切った観測を用いるよりも，ゴールからの距離に応じて区切り方の大きさが変化することが知られている．これらの研究では，本質的に同一な状態をまとめていき，状態と観測が一対一で対応する状況を構築することが目的となる．この場合，設計者がとりあえず必要と思われる入力情報をかき集め，エージェントに学習させることで，不要な入力情報が判明する．学習に必要な情報が獲得できる前提である．

これに対して，学習に必要な情報が獲得できない場合がある．これを不完全知覚問題 [Whitehead 90] と言う．エージェントにとっては，隠れマルコフモデルに基づく環境や，部分観測可能マルコフ決定下（POMDPs）

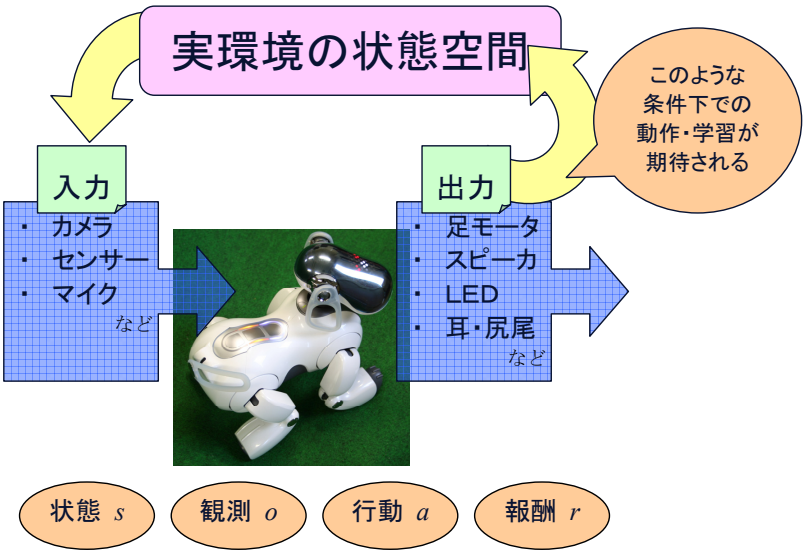

図 1: 強化学習の枠組み

の環境となる．簡単に言うと，最適な行動選択に必要な入力情報を得られない環境である．例えば，迷路において自分の周り数マスしか観測できない場合，異なる複数の場所が同じように見えることがある．しかし，数歩手前のマスから数えながら移動してみると，それぞれが別のマスであると認識できたりする．このように履歴を用いることで区別する方法が提案されているが，どれだけの履歴の長さを必要とするかは問題環境に依存する．また状態遷移が確率的な場合，履歴だけでは一意に状態を決定できなくなる．このときは，区別を諦め，その環境を受け入れる必要がある．代表的な強化学習法である Q-Learning[Watkins 92] は，MDPs 環境下での学習を想定している．その発展系である Sarsa[Rummery 94] や報酬分配型の強化学習法である Profit Sharing[Grefenstette 88, 宮崎 94] は，一部の POMDPs 環境下でも学習を進めることができ [植村 05]，実世界への適用が期待されている．いずれの強化学習も，全ての POMDPs 環境を想定しておらず，MDPs 環境での適用が第一である．そのため，POMDPs 環境への拡張は，おのずと限界が生じる．

そこで本研究では，人工知能の原点に立ち返り，生物が不完全知覚問題にどのように対応しているのかを観察し，今後の方向性を見つけることを目的とする．以下，2 章では，生物が不完全知覚問題に出会ったとき，どのような行動をとったか紹介し，3 章では，それらの行動選択が強化学習から見てどのような政策に従ったものであるのか検討する．そして，4 章でまとめとする．

2. 生物の学習

我が家の話になり恐縮であるが，我が家には猫が一匹いる．人間がルールを決めて接し，猫がそのルールを学習する過程は，人工知能の観点からとても興味深い．猫の餌は，基本的にドライフードを常時置いているが，それとは別に一日一回，ウェットの餌もあげている．このウェットの餌は，かなり好物のようであり，餌付けに最適である．

2.1. 事例 1

当初は昼に一回あげていたが，その後，朝にあげるようしたところ，猫も学習し，餌をねだりに起こしに来るようになった（図 3）．そのうち，猫の方が時間をさかのぼりすぎ，朝の六時，そして五時ごろにねだりに来るようになった．さすがに，早起きができなくなったので,「起きたらあげる」にルールを変えたところ，現在は起きるまで待つようになった．

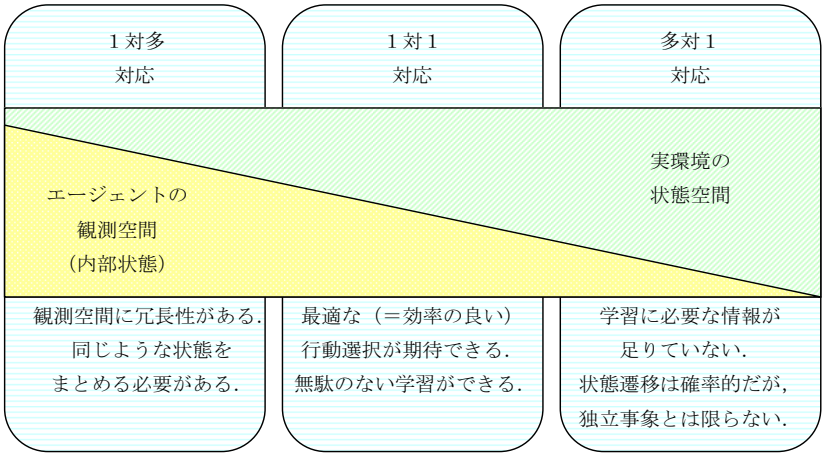

図 2: 状態空間と観測空間の対応

2.2. 事例２

次に，餌をあげる人間を曜日に対して固定した．燃えるゴミの日である火曜日と金曜日は，ゴミ出しのために妻が先に起き，餌をやる．それ以外の曜日は私が先に起き，餌をやる．一週間で見ると，中２日と中３日と，規則性を見つけるには少々困難な条件である．

猫がどのように学習を進めるか観察してみたところ，特に曜日に関しては学習するつもりがないようであり，今のところ「どちらかを起こせば餌がもらえる」という態度である．

2.3. 事例３

その後，引っ越したため，ゴミ出しの時間が朝でなくなった．まさに，ランダムに起きた方が餌をやることとなったが，事例２と同じく起きるほうを起こす行動選択を取っている．ただし最近は，よく餌をくれる方を最初に起こしている傾向がある．記録をつけていないため統計的なデータではないため，今後の課題である．

3. 考察

事例１では，時間に対するパラメータを認識していることがわかる．また，報酬の先読みを行い，自分で時間を判断している．そして，報酬がもらえなくなると，補正をかけて調整を行っている．

事例２では，より長い時間に対するパラメータである．この場合，行動選択を間違えても学習者には罰がないため，手当たり次第行動を選んでも問題がない．そのため，学習しなくても，さほど問題がないと考えられる．

事例３では，学習者にしてみれば，事例２と大きく変わらないため，同じ行動選択になっていると考えられる．

これらのことから，次の二つのことが考えられる．まず，時間に関しては，しっかりと認識しており，必要な場合には入力情報として利用している．ただし，最初から利用しているようではなく，継続して報酬を得られるようになってから，補正として使用している感じである．

次に，行動選択の結果報酬がもらえない場合の扱いである．事例１のように，早起きをしてみたものの餌がもらえない場合は，その行動選択自身が無駄であるため，効率の良い行動選択となるよう学習を進めている．しかし，事例２や３のように，起こす人を間違えても，次に別の人を起こせば報酬が手に入るため，大きな罰ではない．しかも，行動選択を間違えても，次に正しい行動を選択する場報酬がもらえるため，間違えた行動をとっても報酬からさほど遠ざからない．

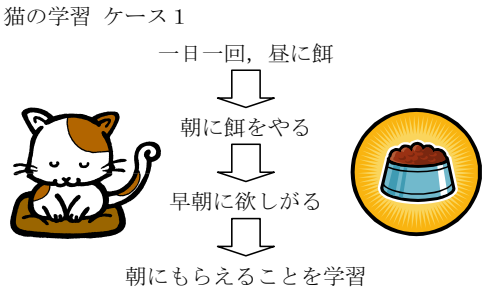

図 3: 生物における不完全知覚問題への対処

強化学習のモデルで考えてみると，時間軸に関しては，観測空間の補正を上手に行っていることが見て取れる．広い観測空間を狭めるのではなく，狭い観測空間から広げていき，必要な広さの空間になるのを見極める流れである．

報酬がもらえない場合の扱いであるが，行動の価値が大きく異なる場合と考えられる．つまり，行動選択を間違えると，報酬から大きく遠のく場合と，あまり遠のかない場合である．後者の場合は，行動選択を間違えても特にダメージではないため，学習を放棄しても問題がないが，前者はダメージが大きいため，しっかりと学習する必要がある．

POMDPs 環境においても，困難な場合の政策として，必要な行動から一つをランダムに選択する方法がある．正答を選ぶために必要なパラメータ（曜日）を知覚できないため，偏りのある行動選択を行うと裏目に出る危険性がある．そのため，ランダム選択が無難な性能を示すこととなる．猫も，この政策に従っていると考えられる．二人のうちどちらを起こすかという選択は，過去の経験を活かせば統計的な有意差を見つけ出すことが可能であるが，そのためにはかなりの労力を費やす．また，その労力を費やして百発百中で当てなくても，何も考えずに行動選択し，ほぼ同じ報酬を獲得することが可能である．コストパフォーマンスを考えると，後者で十分であろう．これは，強化学習の世界で考えると，望ましくない行動選択をした時に，どれがけコストがかかるかを考慮する必要があることを示している．一般に学習の世界では，どの行動選択も目標状態へ（いつかは）至る．そのため目標状態へ近づく行動以外の行動を選択したときは，遠回りをする．つまり，どれだけ遠回りになるのかを示す指標を導入することで，POMDPs 環境，または MDP 環境の困難さを知りえることができるのではないかと考えられる．

参考文献

[Grefenstette 88] Grefenstette, J.J., “Credit Assignment in Rule Discovery Systems Based on Genetic Algorithms”, Machine Learning, Vol.3, pp.225–245 (1988).

[Watkins 92] Watkins, C.J.C.H. and Dayan, P., “Technical Note:Q-Learning”, Machine Learning, Vol.8, pp.279–292 (1992).

[植村 05] 植村 渉, 上野 敦志, 辰巳 昭治, “POMDPs 環境のためのエピソード強化型強化学習法”, 電子情報通信学会論文誌, Vol. J88-A, pp. 761–774, (2005).

[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”，人工知能誌, Vol.9, No.4, pp.580–587 (1994).

[Whitehead 90] Whitehead, S. and Balland, D., “Active perception and reinforcement learning”, in Proc. of the 7th International Conference on Machine Learning, pp. 162–169 (1990)

[Rummery 94] Rummery, G. and Niranjan, M., “On-line Q-learning using connectionist systems”, Technical Report CUED/F-INFENG/TR 166 Engineering Department, Cambridge University (1994)