知能と複雑系　123－24
（2001. 1. 11）

強化学習型情報処理における人間の行動決定について

藤崎 恵美子*　松本 健一*　井上 克郎*,†

*奈良先端科学技術大学院大学 情報科学研究科
〒 630-0101 奈良県 生駒市 高山町 8916-5
†大阪大学 大学院基礎工学研究科
〒 560-8531 大阪府 豊中市 待兼山町 1-3
{emiko-fu, matumoto, k-inoue}@is.aist-nara.ac.jp

あらまし　本研究では，強化学習研究における探索 (exploration) と搾取 (exploitation) のトレードオフ状況において，人間がどのように行動決定を行っているか，そして行動決定に関わる学習要因は何かを探るため実験を行った．その結果，個人により行動方略の違いがあること，また，個人内に「これだけは確保しておきたい」という報酬の最低量の基準の存在が示唆された．学習者は自身の持つ「最低基準量」に現在までの報酬が達しているかどうかを確認しながら，残り行動数を見て方略を決定していると考えられる．

キーワード　探索 (exploration) と搾取 (exploitation) のトレードオフ，行動方略，報酬の最低基準量

Decision-Making of Human Information Processing on Reinforcement Learning Task

Emiko Fujisaki*　Ken-ichi Matsumoto*　Katsuro Inoue*,†

*Graduate School of Information Science Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0101
†Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama-cho, Toyonaka, Osaka 560-8531, Japan
{emiko-fu, matumoto, k-inoue}@is.aist-nara.ac.jp

Abstract　Trade-off between "exploration" and "exploitation" is one of the unsolved problems in the reinforcement learning. In this paper, we experiment on nine subjects with reinforcement learning task to specify the factors with which lerners deside their strategy. As the result of the experiment, strategies of learners can be classified into five groups. In addition, we found three major factors in deciding their action; "the target rewards of the task", "the residual number of actions in the task", and "the current rewards of the task".

key words　trade-off between "exploration" and "exploitation", action strategy,target rewards

1. はじめに

強化学習は機械学習のなかでも，環境と学習器の相互作用がある，試行錯誤により学習をすすめるなど，教師あり学習や教師なし学習とは異なる複雑な性質をもつ．

強化学習研究については未知の問題が多く存在するが，そのうち探索(exploration)と搾取(exploitation)のトレードオフという学習器の行動決定に関わる問題がある．

強化学習において学習器は，行動に対しての報酬の期待値[value]を計算（行動の評価）して行動を決定する．このとき，行動は「探索(exploration)」と「搾取(exploitation)」という2つの概念の重み付けにより表現される．搾取ではvalueの推定値が最大である行動を選択する，探索ではvalueの推定値を改善するために行動する(図1参照)，というものである．

強化学習では「学習終了時に累積報酬量を大きくする」という目的がある．学習器は過去に試した行動の中から，より多くの報酬を生み出す行動を選ぶ（搾取行動）必要がある．一方で，そのために未だ選択していない行動を試す（探索行動）必要がある．現在のため搾取するか，未来のため探索するか，というジレンマにおちいる[1]．どちらかの行動に偏っては学習がうまく進まないので，2つの行動のバランスをどうとるか，が学習の大きなポイントとなる．

現在のところ，この2つの行動のバランスを課題に応じて適切に決定する方法は確立されていない．学習に関わるパラメータ（2つの行動の選択確率）を人間が設定しシミュレーションを行い，うまく学習できればそれが適切なバランスである，という方法により行っている．

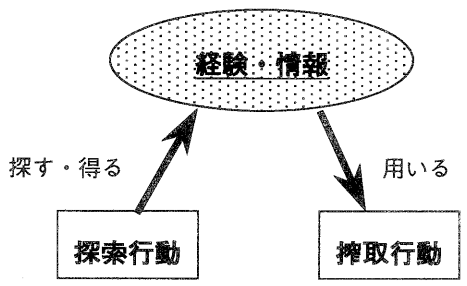

図1：探索と搾取のトレードオフ

一方，人間はこういった強化学習型の情報処理に関わる行動決定を，時間や計算のコストをかけずに逐次行っている．全ての行動を計算してから良いものを選択しているのではなく，その場その場で必要な計算のみを行い要領良く処理しているのである（図2参照）．

本研究では，人間の認知・行動などに関する研究において以前から指摘されているように，人間の振舞いは合理的であるという主張[2]に基づき，強化学習型情報処理について，その認知・行動を含む特性を探る．具体的には，学習者の「探索(exploration)」と「搾取(exploitation)」のバランスに関わる行動決定に必要な学習要素を特定する．また，全ての人間が同様に行動決定を行っているとは考えにくいため，行動方略をいくつかに分類しモデル化を行う．また，データによる方略の特定方法の妥当性を検証する．

本研究の特色は次のとおりである．

(1) 機械学習の一種である強化学習の概念を人間の学習過程に導入する．
(2) 人間の強化学習型情報処理過程をモデル化する．
(3) 人間における探索と搾取のトレードオフの行動決定を明確化する．

2. 関連研究

学習課題における方略に関しては，機械学習，人間の学習の両方から，それぞれ異なったアプローチで研究がなされている．

機械学習の学習方略に関しては，月岡ら[3]が人工知能的観点から分析している．彼らは動的に変化する状況に応じて戦略を自由に使い分ける基準を，学習（主にgreedy method）によって自動的に獲得する手法を提案している．組み合わせ最適化の近似解法を用いた手法で，用意した3つの戦略について比較検討している．だがこの研究では対象領域に制限があり，試行錯誤により学習するという強化学習型の課題ではない．また「より効率のよい解決を行うため」に戦略の使い分けの基準を学習する，という点において，本研究と意図が異なる．

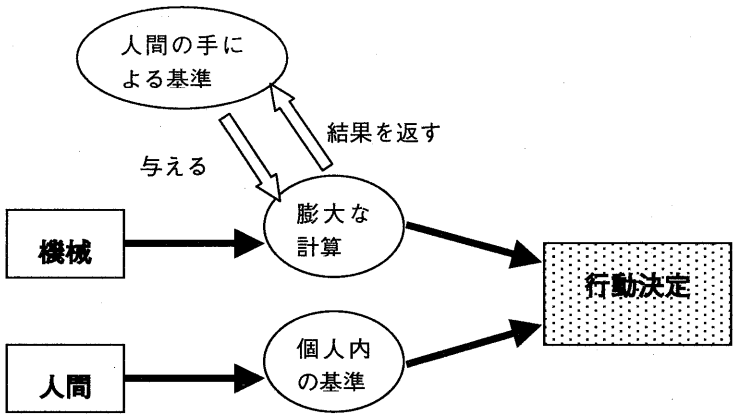

図2：強化学習における機械と人間の行動決定

小堀らは，カードゲームや迷路問題を題材として様々な観点から人間と機械の学習過程を分析している．彼らは迷路問題について，被験者実験の結果から考案した，5 つの探索方略をシミュレーションにより検討している[4]．また，カードゲームを題材に，答えの存在する課題を効率よく解決するための支援システムなどを提案している[5]．ゲームをプレイするプロダクションシステムについて考え[6]，機械学習のアルゴリズムについても模索している．これらの研究は，学習課題の解決過程についてのヒントをもたらすが，認知方略や解決方略等における個人差は特に考慮されていない．

現在のところ，機械学習・人間の学習研究において，解答の存在する問題について「いかに効率良く解答にたどりつけるか」ということや，解答の存在しない問題については機械学習の観点からパラメータを調節して何度もシミュレーションする，といった研究が主である．また，機械学習と人間の学習については，同様な課題についてもそれぞれ異なる分野で，異なるアプローチによって論じられてきた．計算機によるシミュレーションの結果を評価するという心理実験は数多くある．だが，計算機で未だ分かっていない問題について，人間の合理的な（と考えられる）情報処理を分析することで解き明かそうとする研究は少ない．

3. 実験

実験では，性質の異なる 3 つの課題を用意した．これらは以下の条件を満たす．

- 探索行動と搾取行動が（データから）判別可能である．
- 試行錯誤により学習をすすめる．
- 一人で行うことができる．
- ルールが単純で事前知識，知能に関係なく初見で遂行可能である．
- 難易度が適切である．
- 方略に正解が存在せず，自由に行動できる．

課題 1：コマンド入力課題

方向キーに対応した 4 種類の文字からなる正解系列が，あらかじめランダムに設定されている．正解系列は 3 つで，長さは 2～5 である．長さに応じて得点が異なる．被験者は 50 回自由にキーを押し，最終的な累積点数を高くするように求められる．50 回を 1 エピソードとして，10 エピソード行う．事前情報として，正解系列が複数存在し，その長さに応じた得点が入ることが教示されている．

この課題での探索行動は新たなコマンドを適当に入力してみること，搾取行動は過去に報酬が得られたコマンドを入力することである．

課題2：ドア開け課題

3D 迷路の通路にドアがあり，ドアを開けて部屋の中に入るか，開けずにそのまま通路を進むかを選択する．初期状態では，被験者には持ち点が与えられている．ドアを開けるためにはコストがかかる（減点される）が，加点される場合もある．加点される点数はあらかじめ定められた確率によって決まる．ドアを開けるか進むかという選択を 25 回行った時点での得点を大きくすることが目的である．

この課題での探索行動はドアを開けること，搾取行動はドアを開けずに進む（負の報酬を避ける）ことである．

課題3：的当て課題

的が1つだけ表示されたフィールド上で，砲台の角度と強さを調節して的に向けて弾を撃つ．着弾点が的に近ければ近いほど報酬は高い．フィールド上には隠された的が2つあり，それらに当たると高い得点が与えられる．隠れた的を探すか，表示されている的を狙うかは被験者に任されている．弾を 25 回撃った時点で，なるべく多くの累積報酬を得ることが目的である．

この課題での探索行動は，過去に試したことのない組み合わせで，見えない的を探すこと，搾取行動は見えている的，または既に探索行動で発見した得点の高い的を狙うことである．

被験者は大学院生9名で，それぞれの課題について，被験者の選択した行動と得点を記録した．また，実験中は実験者が被験者の行動を観察し，行動の所見を記録した．実験後，方略についての簡単なインタビューを行った．

4. 実験結果

4.1 数値データによる方略特定の妥当性

被験者ごとに次の3つのデータから方略を評価した．

(1)実験後の方略に関するインタビュー
(2)数値データ（方法については 5.2 節を参照）
(3)実験者による，実験中の行動観察

その結果，方略の評価結果は3つの方法間で一致し，数値データによる方略の推定方法の妥当性が示された．

4.2 数値データによる方略の分析方法

各被験者から得られたデータを課題ごと，もしくはエピソードごとに，前半と後半に分け，それぞれの時間帯で方略の判定を行った．9 割以上1種類の行動をとっている場合はその行動をその時間帯での方略とみなした．その結果，方略は次の5種類となった．

a 搾取型：前半・後半ともに搾取行動を一貫して選択する
b 探索型：前半・後半ともに探索行動を一貫して選択する
c 搾取→探索型：前半は搾取行動を主にとっているが後半は探索行動に変化する
d 探索→搾取型：c とは逆に，前半は探索行動，後半は搾取行動に変化する
e 周期型：数回の試行ごとに，探索行動と搾取行動を周期的に繰り返す

方略のデータによる分析結果を表 1 に示す．2 つ以上の課題で同じ方略をとっていたものを網掛けで示してある．ここで示されるように，被験者9名のうち7名が2つ以上の課題において同一の方略をとっている．課題の性質が異なっても，個人内では似た方略をとることが多いと考えられる．

4.3 課題別分析結果

課題1：コマンド入力課題

・ 1エピソード内での方略の変化

1エピソードにつき 50 回の試行中で，a から d の方略がみられた．（表1参照）

・ 課題全体での方略の変化

表1に示すように，1～10 エピソードのうち，エピソードが進むに従い方略が変化した者と，一貫していた者がいた．表中課題 1 の a*，c*，d*はそれぞれ，エピソードごとに a，c，d を繰り返す，という方略である．

図 3，及び図 4 に示すグラフは，試行数の経過による累積得点の推移を表す．搾取行動をとった場合，小さな得点が入り，試行数を重ねるにつれ得点グラフはほぼ直線的な右上がりになる．一方，探索行動では，探索が失敗すると得点が入らず試行数が増えても得点の変化がな

い．探索が成功すると，まとまった試行数に対して大きな得点が入り，得点グラフが階段状になる．

図 3 は搾取型の典型例（被験者 G）である．全てのエピソードについて，得点の入ったコマンド系列をみつけるとそのコマンドを繰り返しているのが分かる．図 4 は探索→搾取型の例（被験者 F）である．1～4 回目では探索行動をとり（4 回目以外は探索失敗とみられる），5～7 回目から後は小さめの得点を短い試行数で何度も得る，搾取志向の傾向がみられる．

課題 2：ドア開け課題

25 試行で，探索行動（ドアを開けること）と搾取行動（ドアを開けずに進むこと）の履歴をとり，試行回数による方略の変化を検討した．

その結果は表 1 の通りである．課題 2 では課題 1，3 に比べ e の周期型の方略をとる者が目だった．

課題 3：的当て課題

・ 2 つのパラメータ調整からみた方略

9 名の被験者の，砲弾の角度と強さという 2 つのパラメータに対する調整行動は，3 種類に分けられた．角度も強さもほぼ一定に調整する方法，角度を固定して強さのみを調整する方法，そして両方とも大きく変化させる方法，である．課題の設定上，角度を一定にすることが可能（試行ごとにリセットされない）である一方で，強さは試行ごとにリセットされることが原因と考えられる．しかし強さをほぼ同じに設定することは可能であるのに，強さのみを同じにする者はいなかった．

次にパラメータ値からみた方略であるが，表 1 にみられるように，a から e の 5 つに分類できた．

・ 得点からみた方略

各人の試行ごとの累積得点をグラフ化して検討した．見えている的，見えていない的に当たったとき，そしてはずれたときにより得点が異なるので，得られた得点により狙った的が推測できる．ここでも方略は a から e の 5 つに分類でき，表 1 で示すように，行動（パラメータ値）からみた方略と得点からみた方略とは，被験者 B と F を除いて一致していた．

表 1：方略のデータによる分析結果

被験者		A	B	C	D	E	F	G	H	I
課題 1	エピソード	a ，b	c	c	a ，b	a	a ，b	a	a	a,b,e
	課題全体	c	c*	c*	d	a*	d	a*	a*	d
課題 2		e	c	b	c	a	e	e	b	b
課題 3	パラメータ調整	e	d	c	d	a	b	b	d	d
	得点	e	b	c	d	a	e	b	d	d

方略の分類は，a:搾取型，b:探索型，c:搾取→探索型，d:探索→搾取型，e:周期型，である．

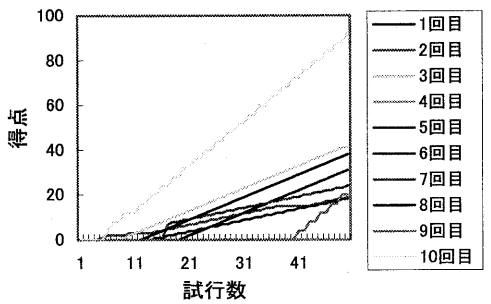

図 3：課題 1 ・搾取型（被験者 G）

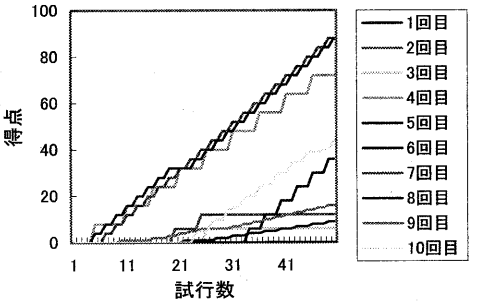

図 4：課題 1 ・探索→搾取型（被験者 F）

5. 行動決定に関わる要因

結果より，強化学習型課題における行動決定に関し重要である要因または情報は，次のものであると考える．

(1) 残り行動数

残りの行動可能回数を考慮して方略を変化させているという報告が目立った．1 エピソードのうち，「いつ報酬が得られたか」が方略決定の大きな要因になっているようである．データによる結果や被験者の報告より，残り試行数に余裕があると探索行動に転じることがある．ただし，探索行動で多くの報酬が得られる系列を発見すると，再び搾取行動に移行する．学習者は残り試行数を適宜モニタしており，残り行動数に余裕がある，すなわち探索行動をとりそれが失敗しても問題ないと判断したとき，探索行動をとると考えられる．

(2) 個人のもつ報酬の「最低基準量」

(3) 現在の得点

得点の推移と被験者による報告，実験者による実験中の方略所見などの結果から，学習者は「少なくともこの程度はとっておきたい」という報酬量の基準を持っていて，それに達した後は残り試行数を見ながら行動するようである．この「最低得点」は個人によって異なり，方略を決定する大きな要因となっていると考えられる．

そうすると，前半と後半で行動を変化させる者について，次のように説明が可能になる．

現在の得点が「最低基準量」に達していなければ搾取を行い，達しているときに残り行動数をみて余裕があるならば探索を行う，という方略．そして逆に，序盤に探索して大きな報酬が得られる行動を探し，残り行動数が少なくなってくると現在の得点をみて終了時までに「最低基準量」を達成しておくべく搾取行動にうつる．

6. おわりに

本研究では，人間の強化学習型情報処理過程での行動決定について，探索と搾取のトレードオフに関する方略とその決定要因を探ること，また実験データにより個人の方略を特定することの妥当性を示すことを目的に実験を行った．実験の結果，実験データや実験者の所見により方略を推定できること，個人により方略の違いが存在すること，そして方略の決定に重要であると考えられる 3 つの要因（残り行動数，個人の報酬に対する「最低基準量」，現在の得点）が示された．今後は，今回の実験で推測された方略を決定する要因について，詳細な検討を重ねていきたい．

参考文献

[1] Richard S. Sutton& Andrew G. Barto：Reinforcement Learning: an introduction I -The Problem-，MIT Press，Cambridge，Massachusetts，pp.3-24，1998.

[2]竹内勇剛，三輪和久：認知資源の合理的利用に基づくモデルのパラメータの推定方法，情報処理学会論文誌，Vol.39，No.7，pp.2124-2133，1998.

[3]月岡陽一，鈴木英之進，志村正道：状況に応じた戦略選択による実時間プランニング，情報処理学会研究報告，Vol.95，No.23，pp.149-156，1995.

[4]小堀聡，小路口心二：迷路探索において利用される情報と知識の検討，情報処理学会第 44 回全国大会論文集，Vol.2，pp.193-194，1992.

[5]中村孝，小堀聡，藤井大輔：問題解決支援のためのアクティブメモ機能について，人工知能学会，ヒューマンインタフェースと認知モデル研究会（第 24 回）資料，SIG-HICG-9403，pp.9-14，1994.

[6]並川青慈，小堀聡，角所収：カードゲームをプレイするプロダクションシステムの学習方法，情報処理学会第 51 回全国大会論文集，Vol.3，pp.171-172，1995.