社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.
ET2000-87 (2000-12)

強化学習型課題の解決における学習者の行動方略と支援

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 学習支援においては，学習過程における人間の認知的・行動的振舞いを明確にしたうえで，学習者の持つ方略や思考，好みなどを把握する必要がある．本研究では，個人の方略を決定する要因を明確にすることで，学習者の行動特性に沿った効果的な支援をめざす． 強化学習型課題の解決行動における「探索(exploration)」と「搾取(exploitation)」のトレードオフに着目し，学習者の方略的な傾向や特徴を明らかにするため実験を行った．その結果，個人により方略の違いがあり，また，「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．学習者は個人のもつ「基準点」に達しているかということと，残り行動数，そして現在の得点をモニタリングしながら方略を決定しているのではないかと考える．これらの情報を用いて，学習者の方略に合わせた効果的な支援について提案する．

キーワード 強化学習，探索(exploration)と搾取(exploitation)，行動方略

Support Systems for Reinforcement Learning Task Forcused on Learner's Action Strategy

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 For supporting human learning, we have to grasp the strategy, thought and tendency of each learner, based on his/her behavior in cognition and action. The aim of this paper is to identify the factors in deciding strategy in reinforcement learning task, from the viewpoint of the trade-off between "exploration" and "exploitation". As the result of our experiment, strategies of learners can be classified into five groups. In addition, we found three major factors in deciding their strategy; "the target rewards of the task", "the residual number of actions in the task", and "the current rewards of the task". In conclusion, we propose the mechanisms of supporting human learning, based on these three factors.

key words reinforcement learning,exploration and exploitation,action strategy

1. はじめに

個人の学習を支援するためには，学習者自身の方略や思考，好みに適応し，それに合わせて振舞うシステムが必要となってくる．そのために，まず人間の学習過程における認知的・行動的振舞いを明確にしたうえで，学習者の方略や思考，偏向などを分析し，それに対し最も適切な情報・行動は何かを探ることが重要である．ここで方略とは，学習や認知などの課題解決過程における行動様式をさす．

本研究では，強化学習型課題における「探索(exploration)」と「搾取(exploitation)」のトレードオフ状況下での人間の行動決定についての実験を通して，課題遂行支援システムを考察する．

2. 強化学習における「探索(exploration)」と「搾取(exploitation)」のトレードオフ問題

2.1 強化学習

強化学習とは，機械学習の一つで，次のような特徴をもつ[1].

- 学習器には「環境」と「行動に対する報酬」が与えられる．
- 環境と学習器の相互作用から学習がすすむ．学習器は環境に行動を起こし，環境はその行動に対する報酬という評価を学習器に与える(図1参照)．学習は知覚・行動・ゴールという3つの面を包含する．
- 有限回の行動後，報酬の累積を最大にすることが学習の目的である．
- 学習器は「どう行動するか」を試行錯誤によって学習する．
- 学習器は方法を学習するのでなく，問題の性質を学びそれに対応する．
- 探索(exploration)と搾取(exploitation)のトレードオフが存在する．この2つのバランスの問題は教師あり学習にはみられない．
- 学習過程において環境と報酬に対する予測とプランニングが重要となる．
- 学習器は行動の効果が完全に予測できない（環境は不確定さを含む）．従って学習器は頻繁に環境をモニタし適切に行動しなくてはならない．

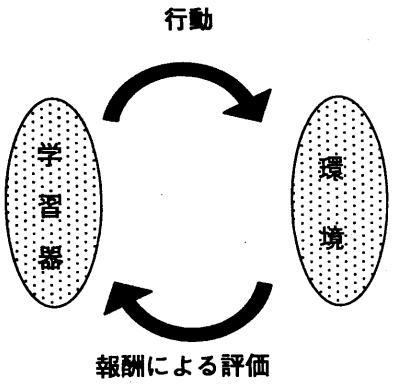

図1：強化学習における環境と学習器の関係

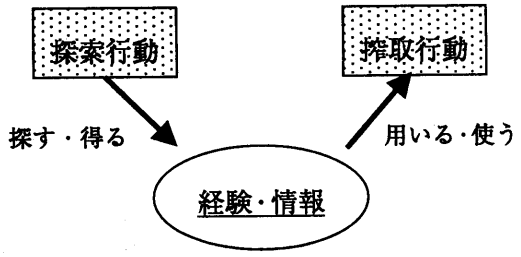

図2：探索と搾取のトレードオフ

2.2 探索(exploration)と搾取(exploitation)のトレードオフ

強化学習において学習器は，過去に試した行動のうち，より多くの報酬を生み出す行動を選ばなくてはならない．だが一方で，新たな情報を得るために以前選択たことのない行動を試す必要もある．(図2参照)．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．この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．

機械学習の研究において，探索と搾取の行動のバランスを決定する方法は現在のところ確立されていない．学習に関わるパラメータを人間が逐一変化させ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，とするものがほとんどである．

3. 関連研究

小堀らは，カードゲームを題材として人間の学習過程を分析し解決支援の手法について提案している[2,3]．これらは「答えの存在する課題をいかに効率よく，間違わずに解決できるか」という観点から，解決過程を分析し分かりやすく効果的なヒントの与え方(機械の支援と協調)や複数ユーザの知識共有などのシステムを提案している．またカードゲームをプレイするプロダクションシステムについても考え，機械学習のアルゴリズムについても模索している[4]．彼らの一連の研究は学習課題の解決支援についてのヒントをもたらすが，認知的方略や解決方略等における個人差には重点がおかれていない．

また小堀らは別の研究で，「正解の経路が 1 通り」で，「迷路の一部のみ見ることが可能」な迷路問題を用いて，人間の探索効率から探索過程を可視化する手法を提案している[5,6]．これらは事前知識が重要でないことや探索過程を検討することについてはカードゲームを題材とした研究と異なる．だが方略に関して解が存在し，それに効率よく到達するため，学習に関わる要素を探る点においては先述の研究と目的を同じくする．

月岡ら[7]は，人工知能的観点から学習方略について分析している．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．これは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．この手法は対象領域が限られているものの，従来手作業で求められていた戦略の選択基準を自動的に獲得する手法として興味深い．しかしこれは計算機による学習をよりよいものとするための手法である．いくつかの戦略を用意してそれらについてシミュレーションを行い評価する，という点において，従来の機械学習研究の域を出ない．

4．実験

4.1 目的

人間の強化学習型課題遂行時の行動決定について，探索行動と搾取行動を中心に，方略の個人による特徴を探る．また，データによる方略の特定方法の妥当性を検証する．

4.2 実験課題

課題の満たすべき条件は次の通りである．

- 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ことができる．
- 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．
- 「正しい方略」というものが存在しない．

以上の条件を満たし，性質の異なる次の 3 つの課題を用意した．

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 回自由に方向キーを押し，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が，正解系列の具体的な個数や長さについては知らされていない．

この課題での 2 つの行動は次の通りである．

- 探索行動：新たなコマンドを適当に入力してみる．
- 搾取行動：過去に報酬が得られたコマンドを入力する．

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．

この課題での行動は次の通りである．
・ 探索行動：ドアを開ける．
・ 搾取行動：ドアを開けずに進む（負の報酬を避ける）．

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．この課題での行動は次の通りである．
・ 探索行動：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
・ 搾取行動：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

被験者は大学院生9名（男性7名，女性2名）で，1名ずつ，3つの課題を順に行った．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．

5．実験結果

5.1 データによる方略特定の妥当性
被験者ごとに次の3つの方法で方略を評価した．
・ 実験後の方略に関するインタビュー
・ データによる方略の推定（方法については5.2節を参照）
・ 実験者による，実験中の行動観察からの方略の推定

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

5.2 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．

5.3 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，試行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．5.2節での分類でいうと，a（搾取型），b（探索型），c（搾取→探索型），d（探索→搾取型），である．しかし，探索型については，探索しようとして失敗したのか，それとも搾取志向だがたまたま報酬を得られる系列を見つけることができなかったのかは，課題の性質（点数の変化から行動を推測する）上，今回のデータだけでは判断できない．

・ 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．

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．

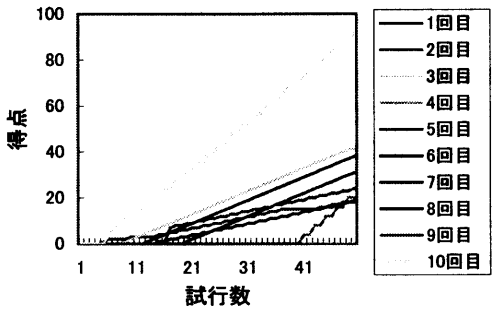

図3：課題1・搾取型（被験者G）

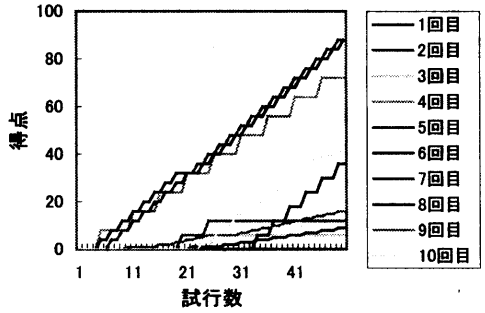

図4：課題1・探索→搾取型（被験者F）

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．1回目から10回目まで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．1～3回目にはほとんど得点が入らず，試行数を重ねても得点は変化しない，また4回目では探索が成功し大きな得点を比較的長い試行数で得ている．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（大きさは同期している傾向がみられる），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，ほぼ次の5つに大別できる.

a)見えている的に対し，角度も強さもほぼ一定に調整する（搾取行動）方略（被験者E）

b) 最初から最後まで大きく変化させる（探索）方略（被験者F，G）

c)前半はほぼ一定で，後半大きく変化させる（搾取→探索）方略（被験者A，C）

d) 前半は大きく変化させ，後半からはほぼ一定範囲で微調整を行う（探索→搾取）方略（被験者B，D）

e)序盤はほぼ一定範囲で，中盤は大きく変化させ，終盤には再び一定範囲内で調整する（搾取→探索→搾取）方略（被験者H，I）

・ 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.

分類すると，探索型（被験者B，G），搾取型（被験者E）搾取→探索（被験者C，F），探索→搾取（被験者D，H，I），搾取→探索→搾取（被験者A），の5つである.

また，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.

6．考察

6.1 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」が方略変化のカギ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.

6.2 学習者に対する支援

今回の実験で推察された，行動を決定する際に重要な3つの要因を効果的に呈示することで，行動決定に対する学習者の負担を軽減し，適切な選択を支援できるであろう.

具体的には，次のような支援を提案する.

- 残り行動数が少なくなってくると警告する
- 個人の行動から報酬の「最低基準量」を推測して，その量に達しない時点で探索に移ろうとすると注意を促す
- 「基準量」を達成するために随時目標までの点数や推定行動数を表示し，最適と考えられる行動を提案する.

7．まとめと今後の課題

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．またこれらのことを用いて，学習者の行動方略の特性に沿った学習支援の具体的方法について考察，提案をした．

今回の実験で示唆された，「最低得点の存在」「残り行動数を考慮した計画性」などの方略決定要因について，より詳細な検討を行うため，実験を計画中である．また，今回の実験条件では，正解・報酬等，被験者の運により課題の難易度が変化してしまう．その結果，個人のもつ本来の方略が変化したとも考えられる．

(1) 正解を固定して難易度が全く同じ条件で比較をする

(2) １つの課題についての試行回数を多くする

などの条件での追実験が必要と思われる．

参考文献

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

[2]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.

[3]藤井大輔，小堀聡，中村孝：マルチエージェントを用いた calculation プレイ支援システムの検討，情報処理学会研究報告，Vol.95，No.23，pp.119 - 126，1995.

[4]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.

[5]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.

[6]小堀聡，迷路問題における人間の探索パターンの解析，第 36 回システム制御情報学会研究発表講演会論文集，pp.281-282，1992.

[7]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.