強化学習によるドライバーの経路選択行動シミュレーションモデル*

Simulation model of drivers' route choice behavior with reinforcement learning

鈴木淳司**・櫻井俊和***・宮城俊彦***
By Atsushi SUZUKI**・Toshikazu SAKURAI***・Toshihiko MIYAGI****

1．はじめに

道路の拡充に伴い複雑化するドライバーの経路選択行動や様々な要因で変化する交通状態など，近年の交通現象は複雑で多様なものとなっている。従来の静的な利用者均衡配分では過去の経験から経路選択を判断するという学習の効果を明示的な形で考慮できないことや連続したリンク・パフォーマンス関数しか扱えないがゆえに複雑な交通現象に対するドライバーの経路選択行動の説明力に疑問が残る．ダイナミックに変動する外的要因に柔軟に対応する交通行動モデルを取り入れた経路選択シミュレーション法が必要になってきている[1)].

本論文の主要な目的はドライバーの交通ネットワーク内のルート選別過程における学習振る舞いを公式化することにある．本研究で提案するアプローチは人工知能での分野の最新研究である強化学習を基本にした形で，個々の運転手が各々の最適ルートを探索する過程を定式化したもので，運転手の個性も表現することができる．この手法は利用者均衡配分から確率均衡配分にわたる色々な均衡状態を実現することができる．しかし，強化学習法では確率配分モデルで扱う分散パラメータを必要とせず，また，ニューラルネットワークで必要とする教師信号もつかわないのが特徴である．

*キーワーズ：配分交通、経路選択、交通行動分析、
** 岐阜大学大学院地域科学研究科
（岐阜県岐阜市柳戸1-1、TEL058-293-6107）
*** 岐阜大学大学院地域科学研究科
（連絡先　同上）
****正会員、工博、岐阜大学地域科学科　教授
（連絡先　同上）

2．強化学習

（１）強化学習とは

強化学習(reinforcement learning)とは，試行錯誤から何をすべきかを（どのようにして状況に基づく動作選択を行うか）を学習する機械学習アルゴリズムである[2)].

ニューラルネットワークなどの多くの機械学習(machine learning)は，入力に対しての「望ましい出力」が用意されており、システムの出力を「望ましい出力」と一致させる学習を行う．このような学習法は教師付き学習(supervised learning)と呼ばれる．

一方，強化学習は行動の結果としての報酬という入力信号のみから学習を行う．教師付き学習のように,「望ましい出力」を与えることはできない．このような学習法は教師なし学習と呼ばれ，問題の出題者がその問題に対する答えを用意しなくても良いという利点がある．

本稿では，学習者をエージェントと呼ぶ．強化学習では，エージェントは与えられた目標を達成するための方法を与えられた環境での相互作用から学習する．

エージェントは環境の状態(state)を観測入力し，その入力に対して可能な行動の集合から，ひとつの行動を選択出力する．行動の結果として報酬を受け取り，次の状態を観測入力する．この繰り返しを行い，学習を進めていく．

エージェントは，与えられた報酬から行動選択を改善し，報酬の最大化を目指す．ある状態において，他の条件が同じであれば，大きな報酬を得られる行動は，より選ばれやすくなる．

すなわち，エージェントは報酬を最大化するた

めの状態から行動の写像を学習する．
この写像のことを政策(policy)と呼ぶ．

本研究では，エピソード型タスクを扱う．エピソード型タスクとは，エージェントの初期状態から終端状態の系列に分解できるタスクである．その系列をエピソードと呼ぶ．

（２）経路選択のための学習プロセス

人の学習プロセスを考えてみると，我々が環境との相互作用を通じて学習しているということが思い浮かぶ．幼児が「はいはい」から「二足歩行」の変化の過程において，そこには教師に相当するものはいないが，その幼児は転ばないように試行錯誤を繰り返し，歩けるようになる．

我々の日々の生活においても，我々自身と環境との相互作用が主要な知識源であることは疑いないところである．車の運転を学習する過程，目的地までの最適な経路を学習する過程においても，このような相互作用が学習に影響を与えていると考えられる．相互作用から学習することは，ほとんど全ての学習理論の根底にある基本的な考え方である．

本研究では，このような学習プロセスを模した強化学習を用いて学習するエージェントを，ドライバーと見立てる．そして，与えられたネットワークを複数のエージェント（ドライバー）に同時に学習させることで，ネットワークの均衡状態を実現させる．

（３）Profit Sharing [3)−4)]

本研究では，強化学習アルゴリズムに Profit Sharing を用いる．Profit Sharing は，代表的な強化学習アルゴリズムであり，それぞれの状態ごとに行動の優先度を学習する．状態 s における行動 a の優先度を $W(s,a)$ と表し，エージェントは優先度 $W(s,a)$ を使って行動を選択する．

Profit Sharing は，エピソードが終了後に，エピソードに含まれる状態行動対に対して，一括に次式のように更新を行う．

$$W(s_t,a_t) \leftarrow W(s_t,a_t) + f(t,R\ ,T)$$

すべての s,a に対して：

$W(s,a)=C$　（C は任意の正の定数）

各エピソードに対して繰り返し：

s を初期化

エピソード中の各ステップに対して繰り返し：

W から導かれる重み付きルーレット選択を用いて，

s での行動 a を選択する

行動 a を取り，報酬 R と次状態 s'を観測する

$s \leftarrow s'$

s が終端状態ならば繰り返しを終了

エピソードに含まれるすべての状態行動対に対して：

$$W(s_t,a_t) \leftarrow W(s_t,a_t) + f(t,r_T,T)$$

図１：Profit Sharing のアルゴリズム

ここで，f は強化関数，T はエピソードが終了した時刻である．強化関数 f として等比減少関数

$$f(t,R_T,T)=\gamma^{T-t-1}R_T\ (0\le\gamma\le 1)$$

が知られており,よく用いられている．ここでγは割引率パラメータ，R_T は時刻 T で得られる報酬である．Profit Sharing では，行動選択に次式で表されるような重み付きルーレット選択を用いる．

$$\Pr(s,a)=\frac{W(s,a)}{\sum_{a'\in A(s)}W(s,a')}$$

ここで，$\Pr(s,a)$ は状態 s で行動 a を選択する確率，$A(s)$ は状態 s で実行可能な行動の集合をあらわす．Profit Sharing では報酬はすべて正の値とされている．

３．交通量配分問題への応用

（１）確率的利用者均衡配分 [5)]

従来，確率的利用者均衡配分モデルの解法の一つにロジットルート選択モデルに基づいた交通量配分法がある.すべてのフローが独立であるとき，このモデルでは起点 r から終点 s を k という経路を使って得られる効用 U_k^{rs} が

$$U_k^{rs} = -\theta c_k^{rs} + \varepsilon_k^{rs} \qquad \forall k, r, s$$

となり，c_k^{rs} は物理的な意味での所要時間をあらわし，θ （>0）はスケールパラメータ，ε_k^{rs} はランダム項でありガンベル分布にしたがっている．

$r-s$ 間の経路 k の選択確率は以下の式であらわされる．

$$P_k^{rs} = \frac{e^{-\theta c_k^{rs}}}{\sum_l e^{-\theta c_l^{rs}}} \qquad \forall k, r, s$$

この確率的利用者均衡配分モデルの効用を考える際,認知所要時間を使う．$r-s$ 間を経路 k で通る時の認知所要時間 C_k^{rs} は

$$C_k^{rs} = c_k^{rs} + \xi_k^{rs} \qquad \forall k, r, s$$

となり上の ξ_k^{rs} を使って認知所要時間は以下の式のように変形することができる．

$$C_k^{rs} = c_k^{rs} - \frac{1}{\theta}\varepsilon^{rs} \qquad \forall k, r, s$$

前述のとおり ε^{rs} はガンベル分布にしたがっており，θ は認知所要時間の分散を表すパラメータを示している．

（2）強化学習問題への適応

エージェントをドライバーとみなし，交通量配分問題をマルチエージェント問題として扱う．マルチエージェント問題とは，複数のエージェントが同じ環境に存在し，学習する問題である．エージェントは自分以外の情報は知ることができないので，自分以外のエージェントも環境として扱ってしまう．このため，自分は同じ行動を繰り返していても，他のエージェントの行動により，その行動が良い行動にも悪い行動にも変わってしまうことが起こりうる．

エージェントに与える報酬には，ある正の定数から出発地から目的地までにかかった時間を引いた値を与えるものとする．Profit Sharing は，報酬は正の値と定義しているため，常に報酬が正の値になるような定数を設定しなければならない．

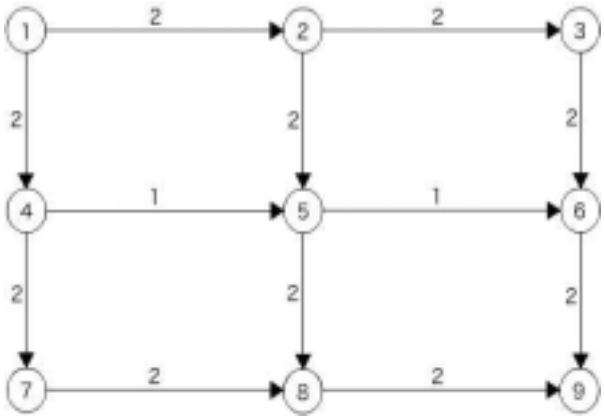

図２：実験に用いたネットワーク

－交通量はノード１から発生し，ノード９を目的地とする．リンク上の数はノード間の移動にかかる時間 t を表す．－

４．実験

（１）問題設定

本手法の有効性を確認するために次のような実験を行った．図２のようなネットワークを設定した [6]．交通量を 100 とし，行動優先度の初期値 $W_0 = 1$，報酬 $R = 10 - \sum t_i$ とした．t_i は i 回目の移動でかかった時間を表す．

（２）実験結果

実験結果を図３に示す．エピソード数（学習の回数）が増えるに従って，エージェントの経路選択が確率的なもの決定的になっていくのが見て取れる．学習の序盤に，Dial アルゴリズムを用いた実験の結果と非常によく似た配分を示している．Dial アルゴリズムを用いた結果を表 1 に示す．

本手法の結果を見ると，従来の確率的均衡配分で用いられる分散パラメータ θ が徐々に減少させていく様子とよく似ていると考えられる．

表１：Dial アルゴリズムの結果

Path	Travel Time	P_k^{rs}	Path Flow
1→2→5→6→9	7	0.196	20
1→4→5→6→9	6	0.534	53
1→2→5→8→9	8	0.072	7
1→4→5→8→9	7	0.197	20

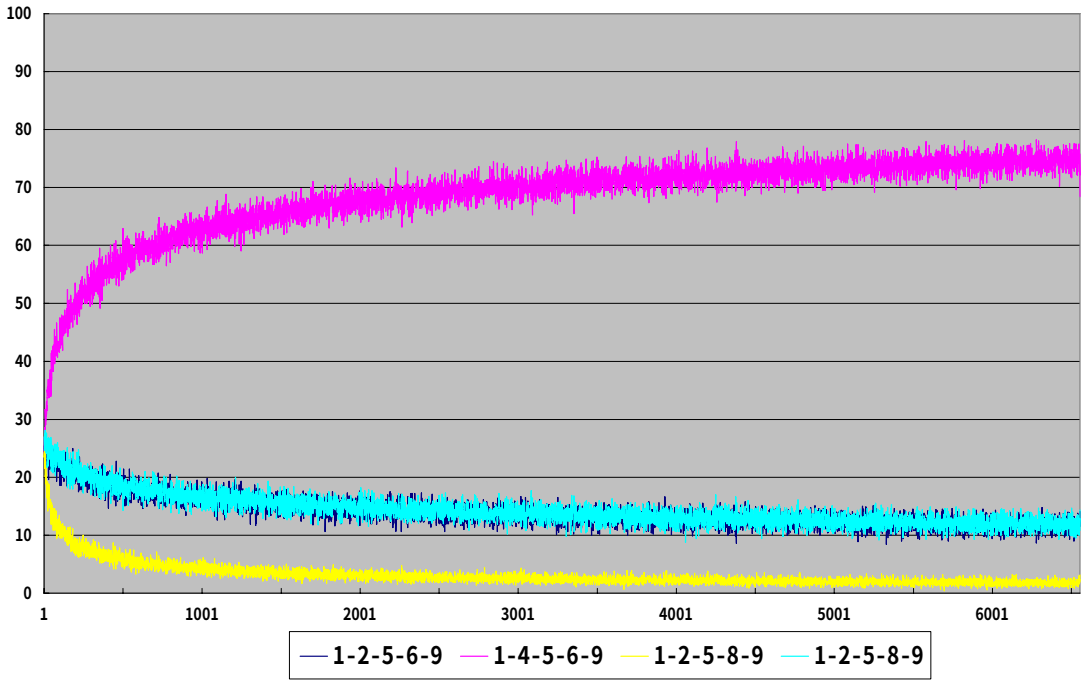

図３：実験結果

－横軸はエピソード数（学習の回数）を示し，縦軸は交通量を示す．
凡例はエージェントが通ったノードの番号である．－

５．考察・まとめ

本研究では，ドライバーの経路選択の学習プロセスを，強化学習を用いて表現した．本研究はまだ始まったばかりで，実装における具体的なモデル化, 理論的な検証などがほとんどできていない．

これから，様々な実験を通して，これらの点を解析していかなくてはならない．特に，本稿で行った実験はリンク間の所要時間は一定であった．従来の交通量配分問題では，リンク間の所要時間は交通量によって変動するのが一般的であるし，現実に近いモデルである．このようなドライバーが他のドライバーからの影響を受けるモデルについても本手法が有効であることを早急に確認したい．

参考文献

１）土木学会：交通ネットワークの均衡分析―最新の理論と解法―，土木学会，1998.

２） Sutton, R. S. and Barto, A. G. : Reinforcement Learning: An Introduction, The MIT Press, 1998［三上，皆川共訳：強化学習，森北出版，2000］

３）松井藤五郎：自律型エージェントの行動学習に関する研究，名古屋工業大学学位審査論文，2003.

４）鈴木淳司，松井藤五郎，世木博久：罰を考慮した Profit Sharing 強化学習法，2003 年度人工知能学会全国大会(第 17 回), 3f4-02, 2003.

５）佐佐木綱，飯田恭敬：交通工学，国民科学社，1992.

６） Yosef Sheffi ： Urban Transportation Networks, PRENTICE-HALL, 1985.