リグレットマッチング型強化学習による経路選択行動分析*

Analysis for route choice behaviour by using a reinforcement learning approach with regret matching model*

宮城俊彦**、石黒雅彦***
By Toshihiko MIYAGI**, Masahiko ISHIGURO***

１．はじめに

交通ネットワーク上の利用者均衡は今までに数理最適化問題や変分不等式として定式化する方法が考案され、研究されてきた。このような利用者均衡配分モデルでは、ドライバーが「自己にとって最善の経路を選択する」という明快な行動原理を基礎に置いている。これらモデルはその論理の明解さ、説得力から実社会に大きく貢献してきた。交通網上のボトルネックを特定し、経路所要時間からネットワークのサービス水準を知り、交通計画作成のための基礎情報を提供してきた。

しかし、これらのモデルではドライバー特性の差異は考慮されず、同じ選好関係、完全情報、完全予見をもって行動する同質なドライバーを仮定している。また、数理最適化問題の定式化、あるいは解の一意性を保証するため、限定されたリンクコスト関数や需要関数が利用されてきた。ITSによる交通情報提供が交通制御や経路誘導において有効に機能するためには、交通情報と人の行動を分析するフレームワークが構築される必要がある。

個々のドライバーの行動記述については交通の分野では非集計選択行動モデルが知られているが、最近のゲーム理論と学習理論の発展は情報と行動の関係をより一般的なフレームワークで分析することを可能にする。

本研究ではドライバーが個人の走行経験から得る交通情報（所要時間に限定する）のみを頼りに経路選択行動を繰り返す場合の選択行動をゲーム論的強化学習によって定式化するとともに、それによって達成される均衡について分析している。本研究ではHart and Mas-Collel によるリグレットマッチング理論を基本にモデルを構成しているが、利用可能な経路の交通情報にドライバーがアクセスできる場合の経路選択行動分析については、既に宮城[1]によって報告されている。本研究は強化学習に焦点を合わせたモデルを提案する。

*キーワード：経路選択、交通行動分析、交通ネットワーク分析
**正員　工博　東北大学教授　大学院情報科学研究科
（〒980-8579　宮城県仙台市青葉区荒巻字青葉6-6-06）
***非会員、工修、東北大学大学院情報科学研究科
（〒980-8579　宮城県仙台市青葉区荒巻字青葉6-6-06）

２．均衡概念

（１）相関均衡

戦略型ゲームにおいて、各プレイヤーはある確率分布に従って生起する同一のシグナルによってその行動を決定する場合、ナッシュ均衡とは違った均衡が現れる。これを相関均衡と呼ぶ。各プレイヤーはシグナルによって互いの行動を調整でき、シグナルの発生に対する期待利得を最大化するように行動する。全プレイヤーがシグナルの生起によって定まる条件付確率qに従って行動しているとき、すべてのプレイヤーiと任意の行動に対して、次式が成立する場合を相関均衡という[2]。

$$\sum_{s^{-i}\in S^{-i}} u^i(s^i,s^{-i})q(\mathbf{s}^{-i}\mid s^i) \geq \sum_{s^{-i}\in S^{-i}} u^i(t^i,s^{-i})q(\mathbf{s}^{-i}\mid s^i),\quad \forall t^i \in S^i \tag{1}$$

ここで$\{u^i\}_{i\in N}$はプレイヤーiの利得関数、$\{S^i\}_{i\in N}$はプレイヤーiの戦略である。式(1)における$q(\mathbf{s}^{-i}\mid s^i)$は他のプレイヤーが自分は$s^i$を選択すると認識している状態で、他のプレイヤーが$\mathbf{s}^{-i}$を選択する確率を示しているので混合戦略である。従って、上式は他のプレイヤーが自分はs^iを選択すると信じている状況で、かつ自分もその選択s^iを変更する動機のない状況とみることができる。(1)の不等式が成立するとき他のプレイヤーの予想$q(\mathbf{s}^{-i}\mid s^i)$は事後的にも満たされ、両プレイヤーの選択に相関が生まれる。上式の両辺に$q(s^i)>0$をかけると

$$\sum_{s^{-i}\in S^{-i}} u^i(s^i,s^{-i})q(\mathbf{s}) \geq \sum_{s^{-i}\in S^{-i}} u^i(t^i,s^{-i})q(\mathbf{s}),\quad \forall t^i \in S^i \tag{2}$$

となり同時確率分布を用いて定義された相関均衡を得る。

（２）Hannan一致性

プレイヤーが他のプレイヤーの選択情報を得ることができない場合には、最適な行動を選択できる保証はなくなる。このような情報が不完備な状況下では、プレイヤーは他のプレイヤーがどのような行動をとろうとも、

ある一定の利得水準が保証されるような行動をとるという行動原理が説得力を持つ。この概念はHannan(1957)によって考えられたことからHannan一致性、または普遍一致性と呼ばれる。Hannan一致性はプレイヤーiが過に得た利得の平均値が、すべての行動について、もしその行動を一貫して行ってきたとした時に実現するだろう利得よりも小さくはならない状態と定義される。

$$\sum_{\mathbf{s}\in\mathbf{S}} u^i(\mathbf{s})q(\mathbf{s}) \geq \max_{t^i\in S^i} \sum_{s^{-i}\in S^{-i}} u^i(t^i,s^{-i})q(s^{-i}) \quad (3)$$

左辺はプレイヤーiが過去において得た利得の平均値、右辺はプレイヤーiが達成しうる期待利得の最大値である。Hannan一致性は相関均衡を含むより広い概念である。

3. モデル

（1）内部リグレット

Hart&Mascollel[2]はリグレットという概念を用いて、内部リグレットを最小化するように行動すれば相関均衡に、また、外部リグレット（Hannanリグレット）の場合にはHannan一致性に至ることを示した。内部リグレットとは「ある行動jの代わりにkという行動をとらなかったことに対する後悔」と定義され、次式で表わされる。

$$D_t^i(j,k) = \frac{1}{t}\sum_{\tau\le t: s_\tau^i=j} \left\{u^i(k,s_\tau^{-i}) - u^i(j,s_\tau^{-i})\right\} \quad (4)$$

プレイヤーはこのリグレットを混合戦略決定の際の基準に用い、正のリグレットが発生する選択肢により大きな確率分布を配分する。リグレットの計算と混合戦略の決定の一連の過程を繰り返していき、リグレットがゼロとなる状況を求めるアルゴリズムをリグレットマッチングと呼ぶ。

リグレットマッチングではプレイヤーは自分の利得構造がわかり、かつ相手の行動を全て観察しており、過去に相手がとった行動に対し、自分が別の行動をとっていたときに実現する利得を計算できるという仮定の上に成り立っている。これはプレイヤーはゲームに参加していて、自分の利得構造を理解して、相手の過去の行動を把握しているという点で、仮想プレイにおける仮定と同様である。Hart&Mascollel[3]リグレットマッチングモデルを改良し、プレイヤーが自分の利得しか知りえず、他のプレイヤーの行動を観測できないという強化学習の仮定で成立するモデルを提案し、この場合にも相関均衡が達成されることを示した。

強化学習モデルの場合、次式で定義される修正内部リグレットを用いる。

$$C_t^i(j,k) = \frac{1}{t}\sum_{\tau\le t: s_\tau^i=k} \frac{q_\tau^i(j)}{q_\tau^i(k)} u^i(k,s_\tau^{-i}) - \frac{1}{t}\sum_{\tau\le t: s_\tau^i=j} u^i(j,s_\tau^{-i}) \quad (5)$$

ここで、$\sum_{\tau\le t: s_\tau^i=j}$とは$\tau\le t$のなかで$s_\tau^i = j$のときだけ足し合わせるという意味である。リグレットの右辺第1項は、実際にはわからない、過去にjのかわりにkをとっていた場合の利得の不偏推定量である。t期の選択が$s_t^i = j$のとき、次期の混合戦略は次式で与える。

$$\begin{cases} q_{t+1}^i(k) = \left(1-\dfrac{\delta^i}{t^{\gamma^i}}\right)\min\left\{\dfrac{\{C_t^i(j,k)\}_+}{\mu^i}, \dfrac{1}{|S^i|-1}\right\} + \dfrac{\delta^i}{t^{\gamma^i}}\dfrac{1}{|S^i|} & if \quad k\neq j \\ q_{t+1}^i(j) = 1 - \sum_{k\in S^i: k\neq j} q_{t+1}^i(k) & else \end{cases} \quad (6)$$

混合戦略の第2項はランダム選択であり、リグレットの比例配分との凸結合を用いることによって、偏りのない探査を可能にしている。δ^i、$\gamma^i \in \left(0, 1/4\right)$はプレイヤーi固有の探査パラメーターである。$\mu^i$は定数の慣性パラメーターで$q$が全て確率分布の範囲に収まるように$\mu^i > 2M^i(|S^i|-1)$を満たすように決定される。$\mu^i$はプレイヤー毎に異なる値を用いる。ここで$M^i = \limsup|u^i|$、$|S^i|$はプレイヤーiの選択可能な戦略数である。

ここで頻度分布z_tを定義する。

$$z_t(\mathbf{s}) = \frac{1}{t}|\{\tau\le t : \mathbf{s}_\tau = \mathbf{s}\}| \quad (7)$$

このとき以下の定理が導かれる。

定理1 (Hart and Mas-Collel,2001)

全てのプレイヤーが修正内部リグレットマッチングに従って行動している場合、頻度分布z_tは確率1で相関均衡に収束する。

（2）Hannanリグレット

Hannanリグレットとは「過去に選択した行動のかわりにkという行動を一貫してとらなかったことに対する後悔」と定義される。

Hart&MascollelはHannanリグレットを用いた場合でも内部リグレットと同様に、プレイヤーは自分の利得しかわからないという強化学習の仮定で成り立つモデルを提案し、Hannan一致性が達成されることを示した。[3]

$$CH_t^i(k)=\frac{1}{t}\sum_{\tau\le t:s_\tau^i=k}\frac{1}{q_t^i(k)}u^i(k,s_\tau^{-i})-\frac{1}{t}\sum_{\tau\le t}u^i(\mathbf{s}_\tau) \quad (8)$$

$$q_{t+1}^i(k)=\left(1-\frac{\delta}{t^\gamma}\right)\frac{\left\{CH_t^i(k)\right\}_+}{\sum_{k'\in S^i}\left\{CH_t^i(k')\right\}_+}+\frac{\delta}{t^\gamma}\frac{1}{|S^i|}$$

$CH_t^i(k)$を修正Hannanリグレットと呼ぶ。$\gamma^i\in\left(0,1/2\right)$は学習パラメーター。上式右辺第1項は過去の選択において常にkを選択していた場合に成立する平均利得の不偏推定量で、第2項は実際に選択した行動の平均利得である。Hannanリグレットに基づいて行動するということは今までの過去の選択の履歴と1つの選択kを比較することである。修正内部リグレットを足し合わせると修正Hannanリグレットを導くことができる。

$$\begin{aligned}\sum_{j\in S_i}C_t^i(j,k)&=\sum_{j\in S^i}\left\{\frac{1}{t}\sum_{\tau\le t:s_\tau^i=k}\frac{q_\tau^i(j)}{q_\tau^i(k)}u^i(k,s_\tau^{-i})-\frac{1}{t}\sum_{\tau\le t:s_\tau^i=j}u^i(j,s_\tau^{-i})\right\}\\&=\frac{1}{t}\sum_{\tau\le t:s_\tau^i=k}\sum_{j\in S^i}\frac{q_\tau^i(j)}{q_\tau^i(k)}u^i(k,s_\tau^{-i})-\frac{1}{t}\sum_{j\in S^i}\sum_{\tau\le t:s_\tau^i=j}u^i(j,s_\tau^{-i})\\&=\frac{1}{t}\sum_{\tau\le t:s_\tau^i=k}\frac{1}{q_t^i(k)}u^i(k,s_\tau^{-i})-\frac{1}{t}\sum_{\tau\le t}u^i(\mathbf{s}^\tau)=CH_t^i(k)\end{aligned} \quad (9)$$

定理２(Hart and Mas-Collel, 2001)

全てのプレイヤーが修正Hannanリグレットマッチングに従って行動している場合、頻度分布z_tは確率1でHannan一致性に収束する。

（３）近視眼的Hannanリグレット

本研究ではHart&Mascollelの修正内部リグレット(相関均衡)、修正Hannanリグレット(Hannan一致性)の他に、新たに近視眼的Hannanリグレットを提案する。

$$\begin{cases}MH_t^i(k)=u^i(k,s_t^{-i})-\tilde{u}^i(\mathbf{s}^t)\\q_{t+1}^i(k)=\left(1-\frac{\delta}{t^\gamma}\right)\frac{\left\{MH_t^i(k)\right\}_+}{\sum_{k'\in S^i}\left\{MH_t^i(k')\right\}_+}+\frac{\delta}{t^\gamma}\frac{1}{|S^i|}\end{cases} \quad (10)$$

Hannan リグレットが過去に選んだ分も含めて行動を評価しているのに対し、近視眼的 Hannan リグレットはその行動が選択された一時点の情報のみを今までの自分の平均値と比較する。リグレットが負になるとき、Hart&Mascollel と同様に式展開を行うと、

$$\sum_{\mathbf{s}\in\mathbf{S}}z(\mathbf{s})\left\{u^i(\mathbf{s}_t)-u^i(\mathbf{s})\right\}\le 0 \quad (11)$$

ここで$z(\mathbf{s})$は頻度分布である。近視眼的 Hannan リグレットは今期の選択の組とこれまでの選択の組の利得の差の頻度分布から見た場合の期待値となっている。つまり全プレイヤーの全行動についての近視眼的 Hannan リグレットが全て非正となった場合、実現している頻度分布のもとではこれ以上その期待値が上昇するような選択の組み合わせが存在しないことを示している。

（４） Q 学習

Leslie&Collins[4]はマルチエージェントシステムにおける繰り返しゲームに Q 学習を応用した。彼らの individual Q-learning は Q ファクターから行動の確率分布を与える混合戦略をロジット型の関数で与えることで、今まで収束が困難であったシャープレイゲーム、N 人マッチングペニーゲームでナッシュ分布に収束することを示した。

本研究では前述の 3 リグレットモデルと individual Q-learning モデル、さらに近視眼的 Hannan リグレットと Q 学習を組み合わせたモデルの 5 モデルを交通ネットワークに適用し、その収束を確かめる。

4. 数値計算例

（１）前提

Braess のパラドクスで有名なひし形の交通ネットワークを対象に（図 1)、３章で示したモデルの収束性と均衡特性を分析する。$n=n_1+n_2$ 人のドライバーを考える(n_1 :OD ペア 1→4 を持つドライバー数、n_2 :OD ペア 3→4 を持つドライバー数)。

リンクコスト関数は線形を仮定（ 図 2)。各ドライバーは与えられた混合戦略に基づき 1 つの経路を純粋選択し、その和として実現する経路交通量から各経路の所要時間を計算する。宮城 [1]は所要時間算出の際、経路交通量に各ドライバーの選択経路の頻度分布を用いているが、本研究ではドライバーの確率選択分布に従う純粋戦略の結果より交通量を求める。これにより、本研究では経路の選択から所要時間の算出までの過程全体を通して粒子モデルとして記述することになり、所要時間のランダム性を内生したモデルとなる。

全てのモデルにおいて、利得はマイナスの所要時間で与える。繰り返し数 50000 回を 1 セットとし、20 セットを行い各セットでの収束を確認した。

均衡確認にはリグレット型モデルではドライバーの各経路の選択回数を基準化した頻度分布とリグレットを、また、Q 学習モデルでは各経路の選択確率と頻度分布を比較する。

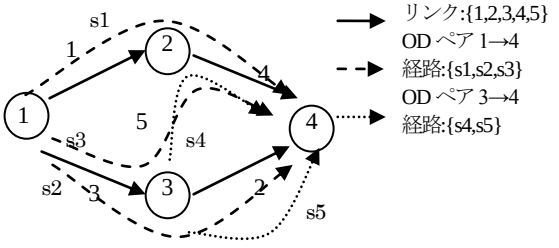

図 1. 交通ネットワーク

表１．リンクフロー関数

$$\begin{cases} 1: t_1 = 50 + x_1 \\ 2: t_2 = 50 + x_2 \\ 3: t_3 = 4x_3 \\ 4: t_4 = 4x_4 \\ 5: t_5 = 10 + x_5 \end{cases}$$

t_i :リンクi所要時間
x_i :リンクi交通量

（２）各モデルの収束状況

単一 OD ペアでナッシュ均衡と等時間均衡が等しくなる場合、異なる場合、複数 OD(2OD)ペアの場合で、それぞれのモデルの収束を確認した。修正内部リグレットモデル、修正 Hannan リグレットモデルでは全ドライバーのリグレットの総和はゼロに漸近し、頻度分布がそれぞれ相関均衡、Hannan 一致性に収束した。相関均衡では毎回のフローパターンは不安定だが、どのドライバーも経路選択の頻度分布は収束し、長期の平均所要時間は全ドライバーで等しくなった。修正 Hannan リグレットモデルでは各セットで頻度分布は収束し安定するが、セットごとを比較しても、収束する頻度分布は一定ではなく、異なっていた。近視眼的 Hannan リグレットモデル、近視眼的 Hannan リグレット+Q 学習モデル、individual Q-learning モデルはどのケースも常に安定的にナッシュ均衡に収束した。

(３)各種設定変更

５ つのモデルのうち、修正内部リグレットモデル、近視眼的 Hannan リグレットモデルで、リンク 5 に料金を付加したケース、さらに時間価値が等しくないケースに適用した。近視眼的 Hannan リグレットモデルでは、リンク 5 の料金が上昇するにつれてリンク 5 の交通量は減少し、ある料金でゼロになった。その料金設定のまま時間価値の高いドライバーを数人設定したところ、時間価値の高いドライバーは料金は高いが所要時間は小さいリンク 5 を使うようになった。一方で修正内部リグレットモデルの場合は、リンク 5 に料金を付加しても頻度分布は変化せず、ネットワークのパフォーマンスは著しく低下した。これが相関均衡による性質なのか、モデリング上の問題なのかはまだわからないが、今後研究を進めて明らかにする必要がある。

5. 終わりに

今回はそれぞれ異なる均衡に収束することが示されているモデルを交通ネットワークに適用し、その収束の違いを見た。相関均衡では毎回のフローパターンは安定しないが平均所要時間は等時間になった。Hannan 一致性では収束はするが、今回はその収束した頻度分布について特徴を見出せなかった。近視眼的 Hannan リグレットモデル、近視眼的 Hannan リグレット+Q 学習モデル、individual Q-learning モデルでは安定的にナッシュ均衡に収束した。今後は各均衡の性質、それぞれの均衡の関連について明らかにしていく必要がある。

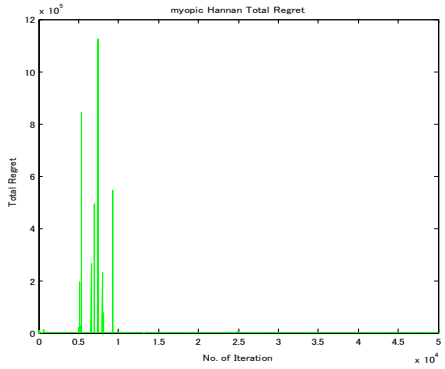

図２　近視眼的 Hannan リグレット・総リグレット減衰曲線

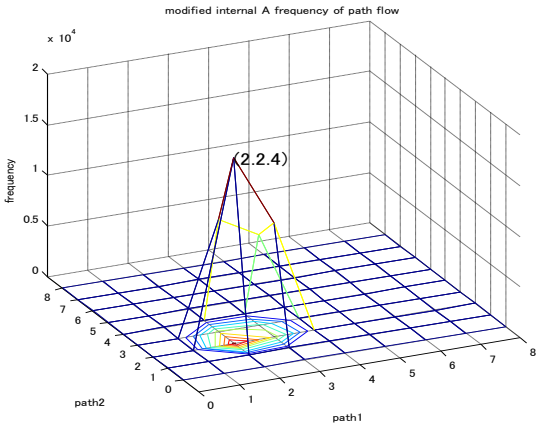

図３　近視眼的 Hannan リグレット・フロー頻度

参考文献

[1]宮城俊彦：経路情報が利用可能な場合におけるリグレット最小化基準に基づく経路選択行動のモデル化, 第 36 回土木計画学研究発表会講演集, 2007.

[2] Hart, S. & A. Mas-Collel : A simple adaptive procedure leading to correlated equilibrium, Econometrica 68(5), pp.1127-1150, 2001.

[3] Hart, S. & A. Mas-Collel : A reinforcement procedure leading to correlated equilibrium, Economic Essays, A Festschrift for Werner Hildenbrand, W.N.G, 2001.

[4] D. S. Leslie & E. J. Collins : Individual Q-learning in normal form games, Siam J. Control optim, 44, 2, pp. 495-514, 2005.