大規模な意思決定問題を解くための学習アルゴリズムの提案*

Learning algorithm for large scale problem of decision-making*

宮城俊彦**・石黒雅彦***

By Toshihiko MIYAGI**・Masahiko ISHIGURO***

１．はじめに

私たちの生活は様々なネットワークの整備によって飛躍的に便利になった。道路ネットワークや情報ネットワーク、エネルギーネットワーク、金融ネットワークなど、私たちのまわりにはネットワークとして表現されるシステムであふれている。ネットワークは今までつながれていなかった人やモノ、情報をつなぎ、仕事や移動といった私たちの日々の活動の範囲や選択肢を拡大してきた。その結果世の中では様々なものがリンクされ、巨大で複雑な多数のネットワークが形成されている。人々はそのネットワークの上で意思決定を行ないながら利益や費用、時間制約、環境の変化などを通じて相互に影響しあい、またネットワーク同士も互いに影響しあう。意思決定問題自体がネットワークを構成している（例：一般の道路網における最短経路、サプライチェーン、物流最適化、ワークシェアリング、テレワークス）。

不確実な環境下での意思決定問題を扱う手法としてはマルコフ意思決定問題（Markov Decision Problem:MDP）があり、多くの研究者によって長年研究されてきた。交通ネットワークの分野にはランダム効用理論を基礎にした確率的利用者均衡配分がある。しかしこれらの研究では環境やエージェントの行動に対して外生的に一応に不確実性を与えており、エージェントの行動の不確実性がどのように他のエージェントの行動に影響するかを見ることはできない。ネットワーク上での意思決定問題においてはリンクで結ばれたノードでの相手エージェントの行動が直接エージェントに影響するため、エージェント間の不確実性の相互の影響を無視できない。

宮城は上述の確率的利用者均衡問題のアプローチとは異なり、交通ネットワーク上での経路選択問題を繰り返しゲーム理論分野の学習理論によって定式化する研究を続けてきた。エージェントたちは交通ネットワーク上で混合戦略に基づくランダム経路選択を行い、よりよい経路を求めて混合戦略を改定する学習行動を行なう。このエージェントたちのランダム選択の結果によって確率的な交通環境を内生的に発生させることができる。このような交通環境の下でもリグレットマッチング型の強化学習アルゴリズムによって利用者均衡に収束することを数値計算によって示した。この一連の研究を進めることによって学習理論を基にしたネットワークゲームと元来の交通理論の整合性をとることが期待されると共に、ネットワークで表現される様々な問題に対して今までの交通理論が積み上げてきたノウハウを応用することができるようになるであろう。しかし今までの学習理論を用いた交通ネットワーク分析は理論研究ということもあり、小規模なネットワークに適用することに終始してきた。今後大規模なネットワークを扱っていくにあたり、選択肢経路が膨大な数になる中でいかにエージェントの選択肢集合を形成するかという課題が生じる。そこで本研究ではErnesto Q.V. Martins and Marta M.B. Pascoalによるｋ最短経路アルゴリズムを利用して、エージェントの選択肢集合を求め、それを基に学習を行なうモデルの提案を行なう。

本研究では意思決定においてとるべき行動が複雑に絡み合っており、ネットワークとして表現されるような問題において、効率的に意思決定ツリーを作成し、コストを最小にするような行動集合を求める手法を提案する。意思決定ツリーとはノードとリンクからなるネットワークで構成され、それぞれノードは意思決定エージェントの状態を、リンクは行動を表わす。このネットワークはスーパーネットワークと呼ばれ、道路ネットワークのような物理的なネットワークを拡張したもので空間的、時間的な広がりを持つスペースタイムネットワークで表わされる。始点と終点が複数ある場合、またはマルチエージェント問題の場合、複数ODペア問題としてこれを解く。ここで意思決定ツリーのリンクコスト（利得）は確率的に変動し、正確にはわからない。またエージェントは探索した経路ツリーの利得しか情報が得られない。

まず２章ではスーパーネットワークによる意思決定問題の表現を行なう。3章ではそのネットワーク上でエー

*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

Tel: 022-795-7495 mail:toshi_miyagi@plan.civil.tohoku.ac.jp）

***非会員、工修、東北大学大学院情報科学研究科

(mail: m-ishiguro@plan.civil.tohoku.ac.jp)

ジェントの経路選択肢集合を生成するｋ最短経路アルゴリズムについて説明する。そして4章においてリグレットを用いた学習理論を説明する。

２．スーパーネットワーク

スーパーネットワークとは意思決定プロセスを可視化し、分析を行なうためのネットワークで Anna Nagurney and JuneDong によって提案された。道路ネットワークなどの物理ネットワークを含み、それを意思決定プロセスの範囲まで空間的、時間的に拡大したものである。スーパーネットワークは道路ネットワークや情報ネットワークといったネットワークを意思決定問題の視点から一般化したネットワーク概念といえる。Anna Nagurney and JuneDong はこのネットワークを用いて意思決定問題を可視化し、変分不等式問題として定式化、数多くの分析を行なっている。本研究ではこのスーパーネットワークを用いて意思決定問題をネットワークとして表現し、学習理論によって分析を行なう。

ノード集合 $N=\{v_1,\cdots v_n\}$ 、有向リンク集合 $A=\{a_1,\cdots a_m\}\in N\times N$ 、$a_k=(v_i,v_j)$ で表現されるネットワーク (N,A) を考える。このネットワーク上での経路は連続するリンクとノードの列として $p=<v'_1,a'_1,v'_2,\cdots,a'_{l-1},v'_l>$ 、$a'_k=(v'_k,v'_{k+1})$ と表わされる。各リンクは行動を表わし、エージェントがそのリンクを選択すると、リンクコスト（あるいは利得）関数が変化する。リンクコスト関数は $l_{a_k}(\mathbf{p})$ で与える。$\mathbf{p}$ は全エージェントの選択経路ベクトルを表す。

３．ｋ最短経路アルゴリズム

（1）表記法

経路集合 P_{st} を起点ノード s から終点ノード t までの全経路の集合とする。部分経路 $sub_{p_{st}}(x,y)$ を $(x,y)\in p_{st}\in P_{st}$ のときの $q\in P_{xy}$ と定義する。リンクコスト $l_{a_k}(\mathbf{p})$ を経路に沿って足し合わせたものを経路コスト $c:\underset{i,j\in N}{\cup}P_{ij}\to\mathbf{R}$ 、$c(p,\mathbf{p}^{-1})=\sum_{a\in p}l_a(p,\mathbf{p}^{-1})$ とする。ここで $\mathbf{p}^{-1}$ は p 以外の全選択経路ベクトル。経路の連結を $p\Diamond q,(p\in P_{ij},q\in P_{jl})$ と表記する。

ここでｋ最短経路集合 $P_K=\{p_1,\cdots,p_K\}$ を定義する。ｋ最短経路集合とは以下の 3 点を満足する経路集合である。

- 全ての経路で閉路を含まない
- $c(p_k)\le c(p_{k+1})$
- $p_k,p_{k+1},\cdots$ の順で決定される

この経路集合を最短経路からの分岐経路を順次求めていくことによって求める。分岐経路とは閉路を含まない経路で、同一の OD ペアを持つ最短経路から分岐ノード $d(q)$ で枝分かれした経路である。

（2）アルゴリズム

最短経路上の各ノードについて分岐経路を計算する。

ステップ１：$k=K$ であれば終了する。$k\le K$ のとき最短経路集合 $P=\{p_1,\cdots,p_k\}$ より p_k を選ぶ。

ステップ２：ネットワーク (N,A) より最短経路 p_k 上の終点 t を除く全てのノード、リンク、分岐ノード $d(p_k)$ に入る全てのリンクを取り除く。

ステップ３：取り除かれたノードのうち、終点ｔに近いノードから分岐ノード $d(p_k)$ までのノードを順に 1 つ $(=v_i^k)$ ネットワークに戻す。

ステップ４：残されたネットワークより v_i^k から終点 t までの最短経路 (v_i^k,t) を計算する。

ステップ５：今までの最短経路の部分経路 $sub_p(s,v_i^k)$ とステップ３で計算された (v_i^k,t) を結合し、新たな経路 $p=sub_{p_k}(s,v_i^k)\Diamond(v_i^k,t)$ を最短経路集合 $P=\{p_1,\cdots,p_k\}$ に含める。ステップ１に戻る。

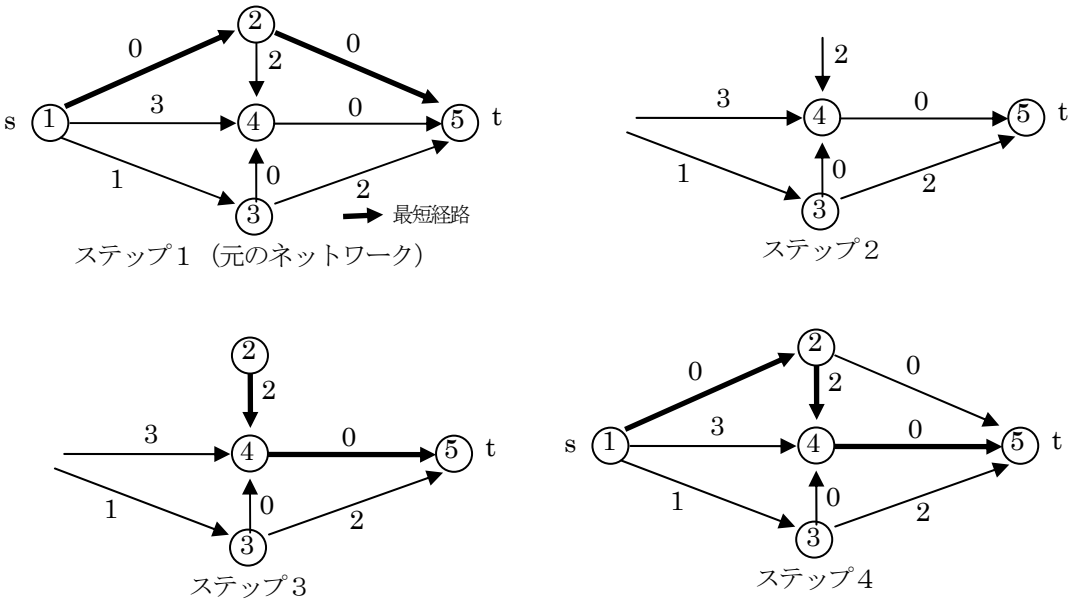

図１　ｋ最短経路アルゴリズム

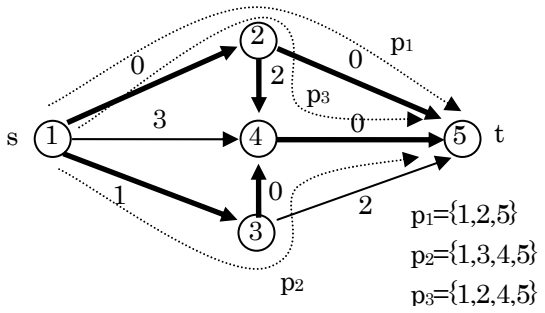

図２　完成したｋ最短経路ツリー(k=3)

４．リグレットマッチング

（1）内部リグレット

Hart&Mascollel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「ある経路jの代わりに経路k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。

$$D_t^i(j,k)=\frac{1}{t}\sum_{\tau\le t:p_\tau^i=j}\left\{c^i(j,\mathbf{p}_\tau^{-i})-c^i(k,\mathbf{p}_\tau^{-i})\right\}$$

ここでp_t^iはエージェントiがt期に選択した経路を表わす。同様に$\mathbf{p}_t^{-i}$はその他のエージェントがt期に選択した経路ベクトルを表わす。

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]の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。

$$C_t^i(j,k)=\frac{1}{t}\sum_{\tau\le t:s_\tau^i=k}\frac{q_\tau^i(j)}{q_\tau^i(k)}c^i(j,\mathbf{p}_\tau^{-i})-\frac{1}{t}\sum_{\tau\le t:s_\tau^i=j}c^i(k,\mathbf{s}_\tau^{-i})$$

ここで、$q_t^i(k)$はエージェントiがt期に採用した経路の混合戦略である。$\sum_{\tau\le t:s_\tau^i=j}$とは$\tau\le t$のなかで$p_\tau^i=j$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$p_t^i=j$のとき、次期の混合戦略は次

式で与える。

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

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$\mu^i > 2M^i(|P_K^i|-1)$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$M^i = \limsup |u^i|$、$|P_K^i|$はプレイヤーiのｋ最短駅路集合の経路の数である。

ここで頻度分布z_tを定義する。

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

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。

$$CH_t^i(k) = -\frac{1}{t}\sum_{\tau \le t : s_\tau^i = k}\frac{1}{q_t^i(k)}c^i(k, \mathbf{p}_\tau^{-i}) + \frac{1}{t}\sum_{\tau \le t} c^i(\mathbf{p}_\tau)$$

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

$CH_t^i(k)$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。

５．計算結果

参考文献

1）T. Miyagi and M. Ishiguro: Modelling of route choice behaviours of car- drivers under imperfect travel information, URBAN TRANSPORT XIV, WIT Press, 2008, pp.551-560.

2）Anna Nagurny and June Dong: Supernetworks Decision-Making for the Information Age, Edward Elgaer Publishing, Northampton, 2002.

3）Ernesto Q.V. Martins and Marta M.B. Pascoal: A new implementation of Yen's ranking loopless paths algorithm, 4OR: A Quarterly Journal of Operations Research, pp.121-133, 2003.

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

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