強化学習に基づく輸送システムの先見性と協調性の獲得

Acquirement of Collaborative and Proactive Behaviors Based on Reinforcement Learning for Transport Systems

向 直人♥ 馮 鈞♦ 渡邊 豊英♠
Naoto MUKAI Jun FENG
Toyohide WATANABE

近年，オン・デマンドな交通システムが注目されている．既存研究の多くは，顧客の要求発生後に，顧客満足度を最大にするような輸送車両や走行経路を割り当てる．一方，本稿では，強化学習手法(Q-学習)を用いることで，都市に潜む輸送要求の傾向を経験から学習し，先見的かつ協調的な運行サービスを提案する．車両は先天的に顧客発生傾向に関する知識を持つ必要はなく，輸送経験から得られる報酬を基に運行経路を獲得する．また，車両は，獲得した顧客発生傾向や，サービス領域のトポロジカルな特徴から各自の担当領域を獲得する．最後に，シミュレーションにより，提案手法の有効性を評価する．

In these years, on-demand transport systems have been focused. Most of traditional studies related to such on-demand systems are regarded as reactive systems. In fact, transport vehicles and their traveling routes are assigned to customers after the occurrences of their demands. In contrast, we propose a proactive and collaborative transport system based on reinforcement learning technique (Q-Learning). Transport vehicles can acquire proactive traveling routes depending on rewards obtained in their transport experiences. Moreover, transport vehicles acquire individual responsible area depending on the topological features of their service area in addition to the rewards of transportations. Finally, we evaluate the efficiency of our transport system by simulation experiments.

1. はじめに

近年，デマンド・バスやインターネット・タクシーと呼ばれる新しい交通システム[1][2]が注目を浴びている．これらのシステムの特徴は，位置情報システム(GPS)を活用して，車両や顧客の位置をリアルタイムに把握することで，顧客にとって最適な車両を割り当てることにある．既存研究の多くは，顧客の要求発生後に，顧客満足度を最大にする車両を割り当てるといった反射的な最適化問題を扱っている．一方，本稿は，強化学習手法(Q-学習)を用いることで，都市に潜む輸送要求の傾向を経験から学習し，先見的かつ協調的な輸送サービスの構築を目標とする．

先見的とは，輸送経験から顧客の発生分布を予測することで，車両が顧客の発生に対して先回り的な行動をとることを意味する．例えば，空車状態において，車両は新たな顧客の発生が見込める位置に事前に移動することで，顧客の待機時間を減少させることができる．また，経路選択において，単に顧客の乗降位置間の最短経路を選択するのではなく，他の顧客の乗降が見込める経路を選択することで，効率的な運営が可能となる．本稿では，このような先見的行動を一般化するために，顧客の乗降を車両の報酬としてとらえ，車両は将来得られる期待報酬が最大となる行動を選択する．また，協調的とは，輸送負荷を車両間で分散することを意味する．本稿では，各車両に担当領域を定め，リレー形式に顧客を輸送することを試みる．サービス領域のトポロジカルな特徴や，学習によって獲得した報酬の期待値に基づき領域を分割（クラスタリング）する．

本論文の構成は以下である．2 章で提案するリレーに基づく輸送システムを形式化する．3 章で顧客発生傾向の Q-学習を用いた獲得方法，4 章でサービス領域のクラスタリング方法を提案する．5 章でシミュレーション実験により提案手法を評価し，6 章でまとめと今後の課題を述べる．

2. 形式化

提案システムを以下のように形式化する．サービス領域 A を式(1)で与える．ノード n は顧客の乗降位置を表し，エッジ e はノード間の経路を表す．ノード数をグラフの大きさとみなし $|A|$ と表す．エッジ長は全て同一であるとする(ノード間の遷移に必要な時間は全て等しい)．また，一方通行や車線数等の制限は考慮しない．

$$\begin{cases} A = (N, E) \\ N = \{n_1, n_2, \cdots\} \\ E = \{e(n, n') \mid n, n' \in N\} \end{cases} \quad (1)$$

顧客の要求発生を以下のようにモデル化する．一般に，顧客の発生分布は一様ではない．出勤時間であれば，住宅街からビジネス街へといった傾向が存在する．そこで，サービス領域内に存在する要求発生傾向を式(2)で与える．フロー f は，乗車位置 n_r，降車位置 n_d，発生確率 η によって特徴付けられる．サービス領域には，特徴の異なるフローが複数重なり合って存在しているものと考える．

$$\begin{cases} F = \{f_1, f_2, \cdots, f_m\} \\ f = (n_r, n_d, \eta) \end{cases} \quad (2)$$

K 台の車両を式(3)で与える．走行速度や最大顧客乗員数は全車両で同一とする(車両の能力は等しい)．よって，担当領域の割当ては車両の能力に依存しないと考える．つまり，車両が分割されたどの領域を担当したとしても，得られる成果は等しい．

$$V = \{v_1, v_2, \cdots, v_K\} \quad (3)$$

ここで，車両の担当領域を考える．サービス領域 A の部分集合を分割領域 SA とする（分割領域間のオーバーラップは

♥ 学生会員 名古屋大学大学院情報科学研究科博士後期課程 naoto@watanabe.ss.is.nagoya-u.ac.jp
♦ 非会員 中国河海大学計算機及び情報工学院 fengjun-cn@vip.sina.com
♠ 正会員 名古屋大学大学院情報科学研究科 watanabe@is.nagoya-u.ac.jp

ない). 分割領域SAに含まれるノード集合を$SA(N)$, エッジ集合を$SA(E)$と表す. また, 接続領域$CA_{SA\to SA'}$は, 条件$e(n,n')\in SA(E), n\in SA(N), n'\in SA'(N)$を満足する, ノード$n$とエッジ$e$の集合とする. 分割領域と接続領域を用いて, 担当領域TA_kを式(4)で定義する.

$$\begin{cases} CA_k = \bigcup_{SA\in A} CA_{SA\to SA_k} \\ TA_k = SA_k \cup CA_k \end{cases} \quad (4)$$

担当領域TA_kは, 自身の独立した領域（分割領域）と, 隣接する担当領域に重複した領域（接続領域）の和で構成される. 車両は担当領域内のみを移動可能であるとする. 担当領域外への輸送要求が発生した場合は, 顧客を接続領域内で一旦降車させ, 隣接する領域の担当車両に委託する.

3. 学習

車両は担当領域内の要求発生分布を強化学習(Q-学習)[3][4]によって学習し, 政策(どの経路を選択すべきか)を獲得する. 本章では, 「状態-行動」, 「報酬」, 「刑罰」の定義を順に述べ, 最後に, 学習した報酬の推定値に基づき, どのように政策を獲得するかを述べる.

3.1 状態ー行動

車両の時刻tにおける状態s_tを, 現在ノードn_tと, そこに至るまでの長さδのノード履歴$(n_{t-\delta},\cdots,n_{t-1})$で与える. よって, 状態$s_t$は式(5)で表される.

$$s_t = (n_{t-\delta},\cdots,n_{t-1},n_t) \quad (5)$$

ノードnに接続するノード集合を$\{L(n)\,|\,e(n,n')\in E\}$と表す. 状態$s_t$における行動は, 現在ノード$n_t$に接続するノード集合$L(n_t)$から選択され, 式(6)で表される.

$$a_t = n_{t+1}\in L(n_t) \quad (6)$$

ここで, 履歴長$\delta=1$のとき, 状態と行動の組は式(7)で与えられる. この状態と行動の組に対して, 報酬の期待値(Q値)を定める.

$$(s_t,a_t) = (n_{t-1},n_t,n_{t+1}) \quad (7)$$

この状態と行動の定義は, 経路選択が現在のノードのみでなく過去にどの経路を辿ったかに依存することを表している. 図1に示すように, たとえ同時刻にノードn_3に到着したとしても, ノードn_1を経由した車両と, ノードn_2を経由した車両とでは, 次の行動によって得られる報酬の期待値が異なる. 結果的に, 履歴長δは, ある輸送要求の乗車位置と降車位置の関連度を表す.

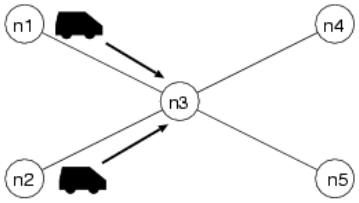


図1 車両の状態と行動
Fig 1. States and Actions of Vehicles

3.2 報酬

車両がノードnに到達したときに得られる報酬を式(8)で与える. ここで, $R_r(n)$はノードnで乗車する顧客の人数であり, ω_rはその重み係数である. また, $R_d(n)$はノードnで降車する顧客の人数であり, ω_dはその重み係数である. 重み係数のバランスは, 車両の輸送方法に影響する. 例えば, 乗車係数ω_rを大きく, 降車係数ω_dを小さくすると, 車両は一度に多くの顧客を乗車させることを好む. 逆に, 乗車係数ω_rを小さく, 降車係数ω_dを大きくすると, 車両は顧客を頻繁に降車させることを好む.

$$R(n) = \omega_r\cdot R_r(n) + \omega_d\cdot R_d(n) \quad (8)$$

3.3 刑罰

車両がノードnに到達すると, ノードnで乗車する顧客の要求は満足される. よって, 直後のノードnで得られる乗車報酬は0となってしまう. そこで, 期待報酬は時間経過に従って本来の値に近付くと考える. つまり, ノードnの経過時間$idle(n)$を考慮して, 期待報酬の推定値$Q(s_t,a_t)$を低く見積もる. ここで, 経過時間$idle(n)$とは, ノードn到達後からの経過時間を表す. 刑罰関数$P(a_t)$は式(9)で与えられ, 推定値$Q(s_t,a_t)$の重み係数となる. また, ζは刑罰の最大回数を表している. 刑罰関数$P(a_t)$は, 時間経過と共に0から1に増加し, 最大回数ζを越えると常に1となる（本来の期待報酬値).

$$P(a_t) = \begin{cases} \dfrac{idle(n_{t+1})}{\zeta} & (idle(n_{t+1}) < \zeta) \\ 1 & (else) \end{cases} \quad (9)$$

3.4 政策

学習した期待報酬の推定値$Q(s_t,a_t)$から, 車両の政策をε-ルーレット手法($0<\varepsilon<1$)で与える. すなわち, 確率εで, 車両はランダムに行動を選択し, 確率$1-\varepsilon$で, 車両は式(10)で示されるルーレット手法で行動a_tを選択する.

$$\Pr(s_t,a_t) = \frac{P(a_t)\cdot Q(s_t,a_t)}{\sum_{a'_t\in L(n_t)} P(a'_t)\cdot Q(s_t,a'_t)} \quad (10)$$

4. 負荷分散

本章では, サービス領域の分割手法（クラスタリング）について述べる. 最初に, 手法の概要を述べ, 次に, クラスタリングの評価基準となる適合関数について述べる.

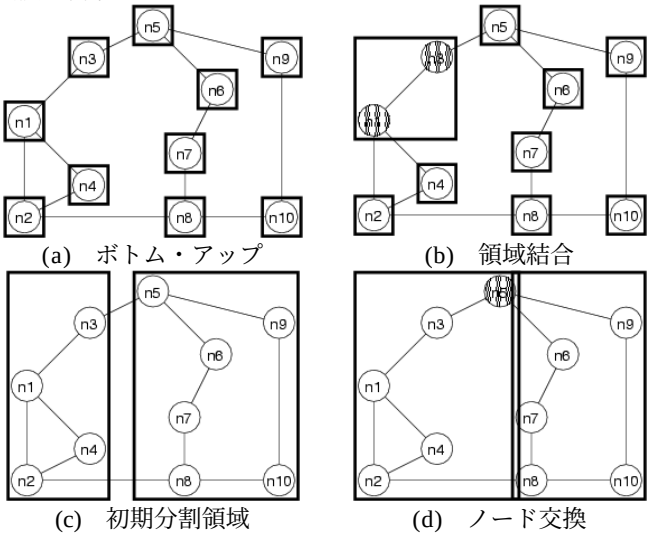
4.1 クラスタリング

クラスタリングにより, サービス領域Aを分割し, その部分集合である分割領域$\{SA_1,SA_2,\cdots,SA_K\}$を生成する. クラスタリングは2ステップ(ボトム・アップ, 最適化)から成る. ここで, 分割領域を評価するために適合関数$fit(A)$を導入する. 適合関数$fit(A)$は, サービス領域Aの分割が, リレ

ー輸送にいかに適しているかを評価する基準となる（その値が小さいほど適している）．適合関数の詳細は次節で述べる．

ボトム・アップでは，分割領域の結合を繰り返すことによって，車両台数Kと同数の分割領域を得る．まず，初期状態では，図2(a)に示すように，ノードを1つだけ含む分割領域を生成する．次に，図2(b)に示すように，分割領域を結合することによって，適合関数$fit(A)$の値が最も小さくなる組を選択し，分割領域の結合を繰り返していく．分割領域の数が，車両台数Kと等しくなったとき，結合を終了する．

最適化では，山登り法に基づき，分割領域間でノードを交換し，適合関数$fit(A)$の値を最小にする．ボトム・アップにより，図2(c)に示すような，車両台数Kと同数の分割領域が得られる．次に，図2(d)に示すように，適合関数$fit(A)$の値がより小さくなるように，分割領域間でノードの交換を繰り返す．ただし，交換するノードは以下の2つの条件を満足する必要がある．分割領域SAからSA'に移されるノードnは接続領域$CA_{SA\to SA'}$に含まれていなければならない．また，ノードnが他の領域に移ることによって，分割領域SAの連結性を破壊してはならない．連結性とは，領域内のいずれのノードからでも，その他の全ノードに辿り着けることを表す．適合関数$fit(A)$の値が改善されなくなったとき，ノードの交換を終了する．



(a) ボトム・アップ　(b) 領域結合　(c) 初期分割領域　(d) ノード交換

図２　サービス領域のクラスタリング

Fig 2. Clustering of service area

4.2　適合関数

クラスタリングの評価基準となる２種類の適合関数（位相関数，報酬関数）を提案する．位相関数はサービス領域のトポロジカルな特徴を考慮する．一方，報酬関数は学習によって獲得した報酬の期待値(Q値)を考慮する．

4.2.1　位相関数

トポロジカルな特徴量として「ノード数平均（式(11)）」，「ノード数標準偏差（式(12)）」，「ノード接続度（式(13)）」の３つを定義する．ノード数平均・標準偏差は，担当領域に含まれるノード数の平均・標準偏差を表す．また，ノード接続度は，接続領域に含まれるノード数の平均を表す．以上の特徴量から，位相関数を式14で定義する．この適合関数を最小にすることによって，分割領域に含まれるノード数が均等化される．また，リレー輸送の中継位置となる接続領域に含まれるノード数が抑えられる．

$$med_t(A)=\frac{1}{|A|}\sum|SA| \tag{11}$$

$$dev_t(A)=\frac{1}{|A|}\sum\sqrt{(|SA|-med_t(A))^2} \tag{12}$$

$$con_t(A)=\frac{1}{||A|^2-|A||}\sum_{SA,SA'}|CA_{SA\to SA'}| \tag{13}$$

$$fit_t(A)=\frac{dev_t(A)+con_t(A)}{med_t(A)} \tag{14}$$

4.2.2　報酬関数

分割領域全体の期待報酬$Q(SA)$を，分割領域に含まれる全てのノードの組合せの期待報酬$Q(s_t,a_t)$の総和とし，式(15)で定義する．

$$Q(SA)=\sum_{n_1,n_2,n_3\in SA(N)}Q(n_1,n_2,n_3) \tag{15}$$

報酬を考慮した特徴量として「報酬平均（式(16)）」，「報酬標準偏差（式(17)）」，「報酬接続度（式(18)）」の３つを定義する．報酬平均・標準偏差は分割領域内で得られる期待報酬の平均・標準偏差を表す．また，報酬接続度は，接続領域で得られる期待報酬の平均を表す．以上の特徴量から，報酬関数を式(19)で定義する．この適合関数を最小にすることによって，分割領域の輸送で得られる期待報酬が最大化される．また，リレー輸送の中継位置となる接続領域で得られる期待報酬が小さくなる．

$$med_e(A)=\frac{1}{|A|}\sum Q(SA) \tag{16}$$

$$dev_e(A)=\frac{1}{|A|}\sum\sqrt{(Q(SA)-med_e(A))^2} \tag{17}$$

$$con_e(A)=\frac{1}{||A|^2-|A||}\sum_{SA,SA'}Q(CA_{SA\to SA'}) \tag{18}$$

$$fit_e(A)=\frac{dev_e(A)+con_e(A)}{med_e(A)} \tag{19}$$

5.　実験

シミュレーション実験の結果を報告する．最初に，1台の車両のみを用いて，学習によって獲得した政策に従って行動する車両と，サービス領域内を最短経路で巡回する車両を比較する．次に，複数台の車両を用いて，領域分割によって車両間でリレー輸送する場合と，領域分割しないで各車両が独立に輸送する場合を比較する．パラメータ設定を表1に示す．

表１　パラメータ

Table 1 Parameters

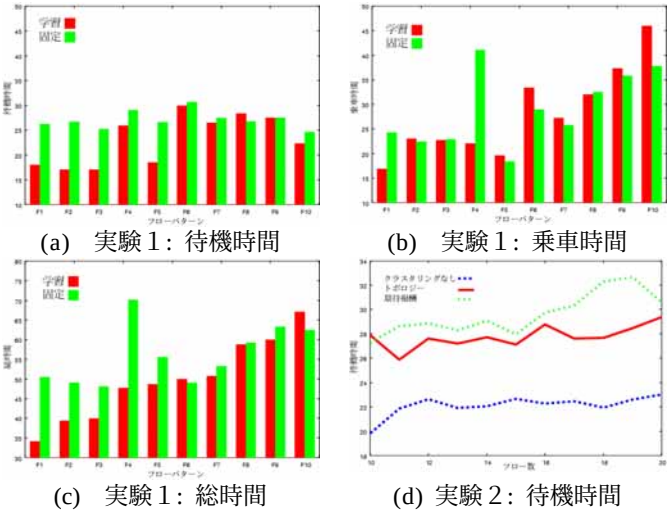
学習率α	0.1	降車係数w_d	1
割引率γ	0.5	刑罰回数ζ	9
履歴長δ	1	ランダム率ε	0.1
乗車係数ω_r	1	総要求発生率$\sum\eta$	0.3

5.1 学習経路 vs. 固定経路

ノード数10, エッジ数15で構成されるサービス領域を生成し, ランダムに10パターンのフロー($F_1, F_2, \cdots, F_{10}$)を発生させた. 図3(a)は顧客の平均待機時間(要求発生から車両が到着するまでの時間), 図3(b)は顧客の平均乗車時間(車両の到着から目的地に到着するまでの時間), 図3(c)は平均総時間をそれぞれ表している. 固定経路では, 待機時間のばらつきが小さくなるに対し, 乗車時間のばらつきが大きくなる. これは, 車両が一定間隔でノードを辿るため, いかなるフローパターンにおいても待機時間は安定するのに対し, 乗車時間はノードを辿る順序に大きく影響を受けてしまうからである. 一方, 学習経路は, フローパターンに合わせて, ノードの到達間隔や順序を変えるため, 待機時間, 乗車時間の両方においてばらつきは大きくなった. また, 総時間に関しては, 10パターン中の8パターンにおいて, 学習経路が優位性を示した. 多くの報酬(顧客の乗降車数)が獲得できる経路を選択することで, 輸送システムの効率が向上するといえる.

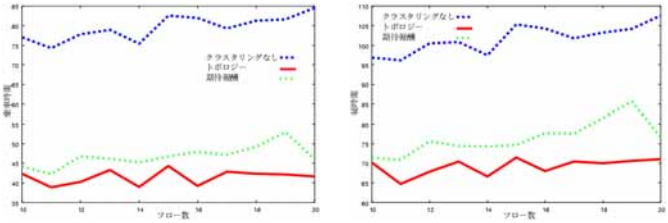
5.2 リレー輸送 vs. 独立輸送

5×5で構成されるグリッド状のサービス領域を生成し, 4台の車両を配置した. フロー数($|F|=10,\cdots,20$)が異なるフロー集合をランダムに発生させた. シミュレーション・サイクルを15000tとし, 最初の5000tを領域全体の学習時間とした. 次に, 獲得した期待報酬値(又はトポロジー)に基づいて領域を分割し, 分割された領域内を5000tで学習させ, 最後の5000tを評価した. 図3(d)は顧客の平均待機時間, 図3(e)は顧客の平均乗車時間, 図3(f)は平均総時間をそれぞれ表している. 領域分割しないで独立輸送する場合, 待機時間は小さくなるのに対し, 乗車時間は大きくなった. これは, 担当領域が広いために, 安定して報酬が見込める乗車報酬ばかりを優先して学習してしまうからである. 一方, 領域分割によるリレー輸送する場合, 両報酬をほぼ平等に学習したため, 待機時間と乗車時間の差は小さい. また, 総時間に関しては, リレー輸送がいずれにおいても独立輸送に勝った. これは, リレーするという負担が発生したとしても, 協調的輸送が効果的であることを示している. また, 位相関数(トポロジー)と, 報酬関数(期待報酬)を比較すると, いずれにおいても, 位相関数が優位を示した. これは, 報酬関数により, 期待報酬を各車両に均等に分散するが, 担当領域の形状が極端にいびつになってしまい, リレー回数が増加してしまうことが原因と考えられる.



(a) 実験1: 待機時間 (b) 実験1: 乗車時間

(c) 実験1: 総時間 (d) 実験2: 待機時間



(e) 実験2: 乗車時間 (f) 実験2: 総時間

図3 実験結果

Fig 3. Experimental Results

6. まとめ

本稿では, 強化学習手法(Q-学習)を用いることで, 従来の考え方とは異なる, 先見的かつ協調的な輸送サービスを提案した. 先見的とは, 輸送経験から顧客の発生分布を予測し, 将来獲得する報酬が大きくなるような経路選択を行うことである. また, 協調的とは, 輸送車両間での効果的な負荷分散を行うことである. 最後に評価実験を行い, 提案手法が輸送システムの効率を向上させることを示した. 今後の課題は, 車両の速度や最大乗車数等の能力差を考慮したクラスタリングを導入することである.

[謝辞]

ご指導頂いた愛知工業大学・石井直宏教授に感謝します.

[文献]

[1] 大田正幸, 篠田孝祐, 野田五十樹, 車谷浩一, 中島秀之: “都市型フルデマンドバスの実用性”, Technical Report 2002-ITS-11-33, 情報処理学会研究報告(2002).

[2] 野田五十樹, 大田正幸, 篠田孝祐, 熊田陽一郎, 中島秀之: “デマンドバスはペイするか?”, Technical Report 2003-ICS-131, 情報処理学会研究報告(2003).

[3] R. S. Sutton and A.G. Barto: “Reinforcement Learning: An Introduction”, MIT Press, Cambridge, MA(1998), A Bradford Book.

[4] H.Santana, G. Ramalho, V. Corruble and B. Ratitch: “Multi-agent patrolling with reinforcement learning”, Proceedings of International Conference on Autonomous Agents and Multi-Agents Systems, pp. 1120-1127(2004).

向 直人 Naoto MUKAI

名古屋大学大学院情報科学研究科博士後期課程在学中. 2003 名古屋工業大学大学院工学研究科博士前期課程修了. 地理情報システムと高度交通情報システムの研究・開発に従事. 情報処理学会学生会員. 日本データベース学会学生会員.

馮 鈞 Jun FENG

中国河海大学計算機及び情報工学院助教授. 1994 中国河海大学コンピュータ学院修士課程修了. 2004 名古屋大学大学院工学研究科博士後期課程修了. 地理情報システム, 空間データベースシステムと空間検索の研究・開発に従事.

渡邉 豊英 Toyohide WATANABE

名古屋大学大学大学院情報科学研究科教授. 1974 京都大学大学院工学研究科修士課程修了. 1975 同大学工学研究科博士課程中退. 工学博士. 統合化環境, 分散協調環境, データベース環境, 教育支援システム, 文書理解に興味を持つ. 情報処理学会, 電子情報通信学会, 日本ソフトウェア科学会, 人工知能学会, システム制御情報学会, 教育システム情報学会, 日本データベース学会会員, ACM, IEEE-CS, AAAI, AACE 各会員.