小型乗合バスシステムにおける最適発車間隔問題のモデル化とその強化学習による獲得手法の提案

蒋 励　　藤田 聡

アジアの各都市で運行されている小型乗合型のバスには、発車タイミングを調整できるメリットと運行者の経験のみにより調整している弱点がある。本論文では、線形経路上に移動する小型乗合バスの運行制御方式をモデル化した。また、オンライン的に変化する不確実性のある利用者状況に対応する発車タイミングを、“試行錯誤”により自律的に獲得するために、POMDP 環境モデルの強化学習の具体的な手法を提案した。さらに、シミュレーション実験を通してその性能を検証した。

Modeling and Proposal of an Approach for Optimal Departure Timing in Xiao-Ba by Reinforcement Learning

Li Jiang and Satoshi Fujita

In this paper, the operation control system of Xiao-Ba, a kind of small size bus which moves on a line, is modeled. In order to obtain the optimal departure timing of Xiao-Ba autonomously corresponding to its conditions, which changes on-line, an approach is proposed by Reinforcement Learning in POMDP environment model. Also, a series of data experiments is executed to evaluate of this method.

１．はじめに

本研究では、アジアの各都市で市民の足として幅広く利用されている小型乗合型のバスシステムに関する効率的な運行方法の提案と評価をおこなう。これらのバスの多くはあらかじめ決められた路線に沿って運行されるが、その発車タイミングを運行者の判断で適応的に変化させることができるという特徴がある。しかし実在する小型乗合バスシステムの発車タイミングは運行者の経験にのみ基づいてなされることが多く、決して合理的とは言えないのが実情である。

本研究の関連研究としては、デマンドバスシステムや AGV システムのための発見的な最適経路選択手法の研究があげられる[1,2]。文献[3,4]では、環境に対する観測が部分的である時のモデル（PODMP）を対象とし、そのようなモデル上で効率的に(最適な行動を獲得するための)強化学習をおこなう方法が提案されている。一方文献[5]では、エレベータシステムの確率的最適制御問題に対する強化学習法が提案されている 。本稿では、その手法を発展させ、エレベータシステムよりも高い自由度と不確実性とをもつ小型乗合バスシステムにおける発車タイミング最適化問題のモデル化と、それに対する強化学習法の提案・評価をおこなう。

本稿の構成は以下の通りである。まず 2 節で線形経路上を移動する小型乗合バスシステムのモデル化をおこなう。提案する強化学習法については 3 節で述べられる。提案手法のよさは 4 節で実験的に評価される。最後に 5 節ではまとめと今後の課題について述べる。

２．モデル化

あるひとつの小型バス路線に注目する。路線の始点から終点までの総延長距離は 10km 程度とし、各バスは始点から終点まで 20 分程度で移動できるものとする。始点と終点の間には 20 個程度の乗降車ポイントが設定されており、各利用者はこれらのポイントからバスに乗降車することができる。バスの運行に関して、本稿では以下のような仮定をおく。

制約条件：バスプール中には十分な数のバスが用意されており、プール中のバスは以下の規則にしたがって順次発車していく：①プール中の先頭車両は、先発車両が出発して一定時間 T(> 0)後には必ず出発しなくてはならない、② 発車は 1 台ずつおこなわれ、同一時刻に複数の車両が同時に出発することはない、③ 各車両の運行速度は同一であり、停車時間は乗降車人数に関わり無く同一である。

始点を出発したバスは、乗降車ポイントを順に通って各ポイントで待っている利用者を拾っていく。利用者は乗車ポイントにランダムに到着し、やってきたバスに先着順に乗車する。ここで乗車拒否は乗客・運転手のいずれの側からもしてはならないものとする。バスの定員を有限値 C とする。あるポイントでバスに乗車した利用者は、終点までに通過するポイントのいずれか（終点を含む）でバスを降車する。以下の議論では、バス運賃は乗車距離によらない定額であるとし、運転手は 1 回の運行で終点に到達するまでに乗車させた延べ人数に比例した収益を得るものとする。

以上ような状況下では、各運転手が実質的におこなえる選択は、バスプール内にいる各時刻で「どれぐらい時間を経てから出発するか」を決定するという一点に集約される。本研究では、各運転手がそれぞれの状況で最適な発車タイミングを獲得する問題を考える。

情報の獲得方法：各乗降車ポイントにはチケットベンダが設置されており、乗車を希望する利用者はあらかじめベンダでチケットを購入の上、そのポイントの

広島大学大学院工学研究科・Graduate School of Engineering, Hiroshima University

待ち行列に並ぶものとする。チケットベンダはプールに設置された端末にオンライン接続されており、プールで待機している各運転手は、端末を通して、どの乗車ポイントでいつ何枚のチケットが購入されたかをすべて把握することができる。また終点には、運行を終了した運転手が運行途中で発生した積み残し数と空席数を記録するための端末が用意されており、これによって、それ以降に始点を出発する運転手に対して情報をフィードバックできるものとする。

評価関数：本稿では、バスの運行効率をバスの運行中の最大乗車率として定義する。いっぽう利用者の満足度は、乗降車後に線路上積み残した延べ人数によって評価し、積み残した延べ人数がゼロのとき、その運行における利用者の満足度は最大になるとする。本研究ではバスの効率を高めると共に利用者の満足度を最大化することを目標関数とする。

３．提案手法

最適発車タイミングを学習する主体のことを以下ではエージェントと呼ぶことにする。エージェントが選択すべき行動はその観測時点からどれぐらい時間を経てから出発するかを決めることであり、そのような行動の集合を以下ではAと記すことにする。強化学習では、エージェントが選択した行動に対して環境が報酬を与えることで、状況に応じた振る舞いの自律的な獲得が実現される。但しエージェントが最適化しようとするのは得られる報酬そのものではなく、自身が保持する行動価値関数（どの状態にいる時にどの行動を選択するのが最も適切なのかを表す関数）の本来あるべきかたちとの距離の最小化である。本稿では、報酬の根拠として、積み残し延べ人数と運行効率を用いる。具体的には、報酬は以下のように与えられる：

積み残し関数 P： $P=\sum_{i=1}^{k}\beta_i p_i \quad (i=1, 2, \ldots, k)$

ここで p_i はポイント i における積み残し人数であり、β_i $(0<\beta_i \le 1)$ は重み係数である

欠員関数 L： $L=(1-\omega)\times C$ （ω：最大乗車率）

報酬関数 R：

	P = 0	0 < P <= Pr	Pr < P
L = 0	Pr - P	Pr - P	0
L > 0	-L	—	—

行動価値関数：行動価値関数の更新は、Q 学習[6]で提案されている方法に習って以下のようにおこなう：

$$Q(s,a)\leftarrow(1-\alpha)Q(s,a)+\alpha\{R+\gamma^{b+T_1}\max_{b\in A}Q(s',b)\}$$

ここで学習率αをうまく設定することにより、学習で獲得される行動価値関数Qを最適なQ値に漸近させることができることが知られている[7]。

ボルツマン選択法[7]：観測された状態における行動選択は以下の規則に従っておこなわれる：

$$P(s,a)=\frac{\exp(Q(s,a)/\tau)}{\sum_{b\in A}\exp(Q(s,b)/\tau)} \quad （\tau：温度係数）$$

一般に POMDP モデル上で最適な政策を多項式時間で得ることは不可能であると考えられており、したがって獲得される政策は自ずと近似的なものにならざるを得ない[4]。POMDP 上での（近似的な）学習を効率のよいものにするためには、状態空間の広さを適切なサイズでおさえることが必要となる。本論文では、状態空間のサイズをおさえるため、積み残し情報を陽に考慮し、利用者到着状況の履歴を増減変化の履歴、すなわち直前のステップに比べてそのステップで到着した利用者数が多くなったか少なくなったかをあらわす１ビットのみで表現するという方法を提案した。

具体的には状態信号を次の３部分により構成する。

① 観測時点に得た利用者待ち行列の到着時刻情報：

$$S_m=\sum_{i=1}^{k}\delta^{k-i}m_i \qquad (i=1,2,\ldots,k)$$

ここで m_i $(i=1, 2, \ldots, k)$ は各ポイントに到着する利用者数であり、定数 δ $(0<\delta\le 1)$ はポイントに与えられる重み係数である（始点から離れれば離れるほど乗車できる確率が下がることを考慮してある）。

② T_1 ステップ前までに出発したバスに関する積み残し関数の値：

$$P=\sum_{i=1}^{k}\beta_i p_i \quad (i=1, 2, \ldots, k)$$

③ 観測時点までの過去 T_1 ステップの間における利用者待ち行列の到着状況の増減変化に関する履歴 S_h：あきらかに S_h の可能なパターン数は 2^{T1} 通りである。

実際に強化学習をおこなう過程では、よくあらわれる状態とそうでない状態とが存在する。これらの状態群をうまく分類することによって、状態の縮約により生じる不確実性を抑えながら、実質的にモデル化する状態数を削減することができる。本研究では、S_m と P の平方根を使って状態をあらわし、状態の数を $\sqrt{S_m}\times\sqrt{P}\times S_h$ にする。

４．シミュレーション実験

実験で想定した利用者の到着状況は以下の通りである。①線路全体に対する到着延べ人数は周期的に変化する。１周期中にあらわれる互いに異なるタイムゾーンの数を T_2 個とする。②各タイムゾーンは U ステップから構成され、各ステップにおいて利用者到着延べ人数はランダムに与えられる。③１ステップ内に各ポイントに到着する利用者数は、ポイントごとに設定さ

れた到着割合に従って与えられる。④あるポイントから乗車した利用者は終点までに通過するポイントの中から等確率でランダムに降車ポイントを選択し降車する。

次に実験で用いたパラメータは以下の通りである。C=20, k=5, T_2=10, U=20 とする。各タイムゾーンで発生する(平均)延べ利用者数は表 1 のように設定される。タイムゾーンごとに到着が期待される利用者のうち、何パーセントずつがそれぞれの乗車ポイントに割り振られるかについては、表 2 に示される５つのパターンのそれぞれについて評価することにする。以下では T=10 とする。したがって、取りうる行動の個数は 10 通りとなる。バスが隣接するポイント間を移動し乗降車させるのにかかる時間を 1 ステップとする。したがって T_1 の値は 5 になる。シミュレーションの 1 試行は 160 万ステップの運行に対しておこなわれる。学習で用いられる各パラメータの初期値はそれぞれ以下のように与えられる：P_r＝30，Q＝0.1，p(s,10)=1，p(s,1)=p(s,2)=…=p(s,9)=0，γ=0.7，δ=0.9。実験に使用した計算機のスペックはCPU:2.4GHz、Memory: 1GBである。

表１　各タイムゾーンに発生する利用者延べ人数

1	2	3	4	5	6	7	8	9	10
100	130	180	180	140	110	80	60	60	80

（*上段：タイムゾーン；下段：人数）

表２　v_i (i=1, 2, 3, 4, 5)までの到着割合(%)

Point / Case	1	2	3	4	5
1	34	10	30	18	8
2	10	34	30	18	8
3	8	10	34	30	18
4	18	8	10	34	30
5	30	18	8	10	34

実験結果と考察

以下では学習の効果を実験終了前の連続する運行 10000 回の平均積み残し延べ人数 P_{av} と平均欠員数 L_{av} によって評価する（これらの値は各々10 回試行で取ったデータの平均値である)。バスの効率と利用者の満足度の間には一般にトレードオフの関係があるが、ここでは報酬関数を利用者の満足度よりもバスの効率を重視するように設定しているので、得られる結果は、L が P に比べてよりゼロに近づく傾向が現れている。

１）学習率αによる影響

図 1 にα＝0.6、図 2 にα＝0.2 の場合の結果をそれぞれ示す（ここではケース 1 についての結果のみ示しているが、ほかのすべてのケースでも収束の様子は同様の傾向が現れる)。αが高い時、最後まで収束の傾向が顕著ではないが、αが下がると学習の収束が明白にあらわれ、さらに下げると収束の様子はゆっくりと悪くなることがわかった。それぞれのケースによって最良なα値は多少違うが、最適値はおおよそ 0.1 から 0.2 の間に存在している。POMDP 環境における強化学習では、与えられる報酬は状態 s_i と行動 a_i のみに依存しているわけではなく、その値の根拠には不確実性が残っている。にもかかわらず Q 値の更新の根拠に報酬をおくのであれば、αを報酬の不確実性の悪影響が抑える程度に小さくすべきであり、この実験結果はそのことを裏付ける形になっている。

２）温度係数τの影響

τを高くすると、エージェントに早めの時点でなるべくランダムに行動を選択し学習することで、τを下げて貪欲的な振る舞いをはじめたときに良い収束傾向が現れた。逆にランダムな選択期間が短すぎると、収束しにくくなる傾向があることもわかった（図 3, 4)。これらのことは、強化学習における探索と知識利用とのバランスをうまくとることが、学習速度を向上させることにつながるということを裏付けている。

３）重み係数β_iによる影響

積み残し関数 P における重み係数β_iは、乗車ポイント i における積み残しをより重視する働きがある。考察の対象とした５つのケースすべてに対して、到着割合が小さいポイントに重きをおくことにより、最大乗車率を高めると共に積み残し延べ人数をある程度に抑えられることが明らかになった（図 5, 6)。状態空間を圧縮するために各ポイントに関する到着利用者状況を S_m 関数で表現したが、これにより、各ポイントへの到着分布の差が大きいとき(本論文では３倍の差がある) S_m 関数に内在する不確実性が大きくなり、その結果、β_iによる調整が不可欠になると考えられる。

今回のシミュレーションでは、最良の結果でも、学習結果が収束する傾向は観測されなかった。これは観測が不完全であることが学習性能を大きく悪化させているためだと考えられる。しかしシミュレーションに使うことのできるメモリ量には制限があるため、今後は不確実さを減らすことのできるような状態空間の圧縮法を提案していくことが不可欠である考えられる。

５．まとめ

本論文では、オンライン的に変化し、不確実性のある小型乗合バスの利用者状況に対して、POMDP 強化学習による状態信号を構築し、最適発車タイミングを自律的に獲得する手法を提案した。さらに実験的にその手法の性能を検証した。今後の課題として学習結果の正確さを高めることなどがあげられる。

参考文献

[1] 内村圭一, 斎藤隆司 and Hiro Takahashi: 遺伝的アルゴリズムによる乗客輸送の最適化, 電学論(D), 117-D(7), 891-897, 1997.

[2] 山根毅史：自動搬送車の動作計画問題に関する研究 http://www.jaist.ac.jp/library/thesis/is-master-2000/paper/tyamane/paper.pdf

[3] W. S. Lovejoy: A Survey of Algorithmic Methods for Partially Observed Markov Decision Processes, Annals of Operations Research 28, 47-65, 1991.

[4] 木村　元,　L. P. Kaelbling: 部分観測マルコフ決定過程下の強化学習, 人工知能学会誌, 12(6):822-830, 1997.

[5] Robert H. Crites and Andrew G. Barto: Elevator Group Control Using Multiple Reinforcement Learning Agent, http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/c/Crites:Robert_H=.html 1998.

[6] Watkins,C.J.C.H and Dayan,P.: Technical Note: Q-Learning, Machine Learning 8, pp.279-292, 1992.

[7] A. G. Barto and R. S. Sutton: Reinforcement Learning, MIT Press, 1998.

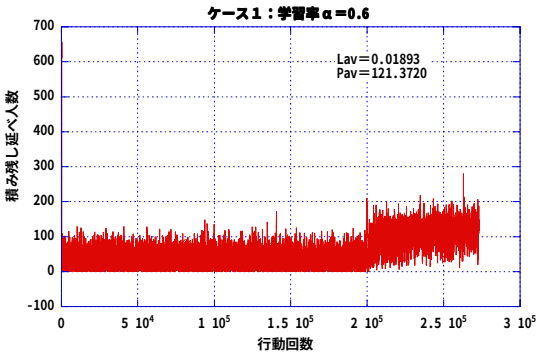

図１　αによる影響（１）

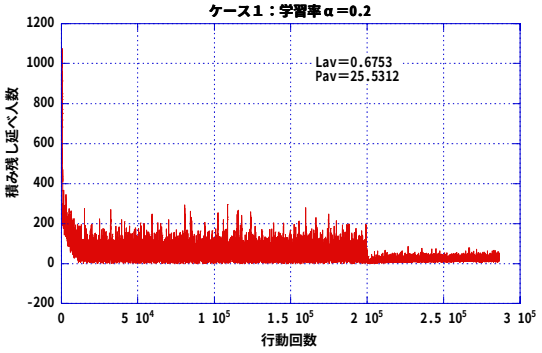

図２　αによる影響（２）

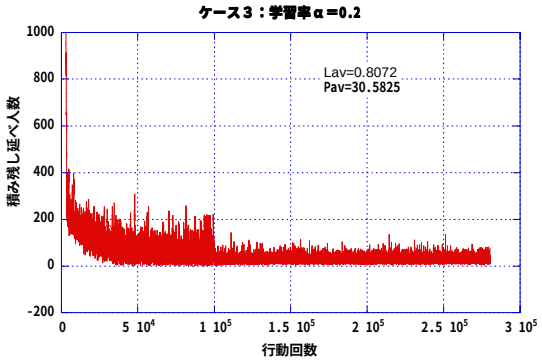

図３　τによる影響（１）

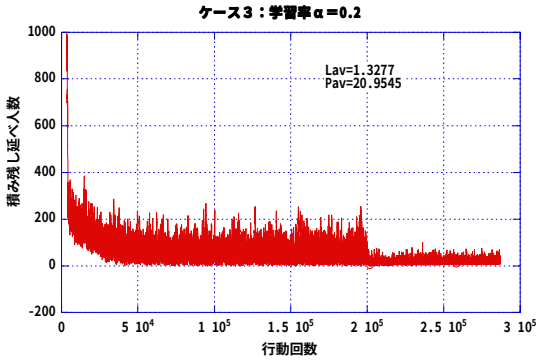

図４　τによる影響（２）

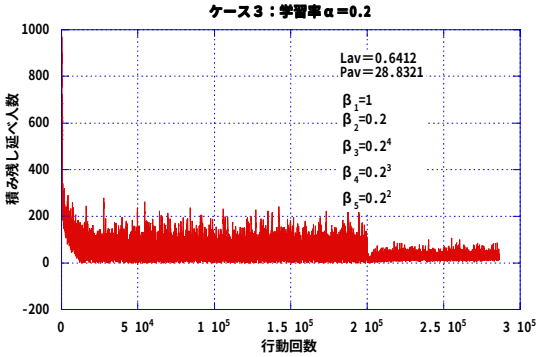

図５　β_iによる影響（１）

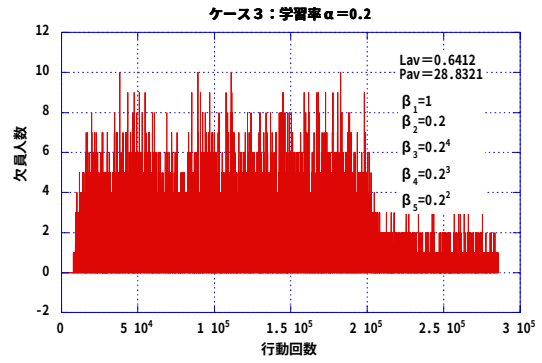

図６　β_iによる影響（２）