Information Processing Society of Japan

Vol. 49 No. 1 情報処理学会論文誌 Jan. 2008

自己代理強化学習による交通信号制御

参沢匡将[†1] 阿部孝司[†2]
下川哲矢[†1] 木村春彦[†3]

現代社会において，交通流の増大による渋滞の発生や事故が増大している．この改善策として信号機による制御が行われている．さらに，近年ではITS（高度道路交通システム）やハイテク信号機の導入などが行われている．また，大規模で複雑なシステムを構築する手法としてマルチエージェントシステムが用いられることが多く，交通信号制御への応用も行われている．しかし，マルチエージェントシステムを適用する場合には，学習に多くの時間がかかるという問題が生じる．本論文では，学習時間を短縮する手法として自己代理強化学習を提案する．これを交通信号制御に適用し，その特性を検証する．

A Traffic Signal Control Using Self Vicarious Reinforcement Learning

Tadanobu Misawa,[†1] Koji Abe,[†2] Tetsuya Shimokawa[†1]
and Haruhiko Kimura[†3]

As improvements for increase of traffic jams and accidents, ITS (Intelligent Transport System) and high-tech signals have been introduced. On the other hands, when such a large system as a traffic control system is designed, multi-agent system is often applied. However, multi-agent systems have a problem of demanding large learning time. In this paper, we present self vicarious reinforcement learning as a method for shortening the learning time and apply in traffic signal control. In addition, we simulate the proposed system using the self vicarious reinforcement learning for traffic signal control and show the usefulness of the self vicarious reinforcement learning.

1. まえがき

現代社会において，車両の保有台数が増大し，これにともない交通流も増大している．この対策として，信号機群による制御があげられる．信号機は，交差点において車両の交通を捌き，交通の安全と円滑を保つために重要な施設である．特に交差点では車両相互の分流，合流，交差など，交通が交錯することから必然的に交通現象も複雑になり，道路上で最も事故や交通渋滞の発生のおそれのある場所である．この交差点において，異なる交通流をできる限り単純化して事故の発生を防止し，また，交通量に応じて通行時間を適切に定めることによって円滑な通行をはかることを目的として信号機が設置されている[1)-3)]．しかしながら，信号機の設置場所・運用方法が適切でないと，かえって事故の増加，交通渋滞の原因ともなることから，その設置・運用にあたっては，慎重を欠いてはならない．近年では交通渋滞に頭を悩ませる警官がIBMのHELP DESKに相談し，渋滞が緩和するCM[4)]が流れるなど関心が高いことが分かる．また，実際に国土交通省道路局の高度道路交通システム（ITS：Inteligent Transport Systems)[5)]や警視庁のハイテク信号機に代表されるように本格的に研究，開発が行われている．

上記ハイテク信号機はLANで接続された信号機が各信号機に設置されたセンサ情報を交換することで通過する車両の台数に応じて通行時間を適時変更する方法である．これはセンサによって環境を観測し，その環境に応じて動作するものとして定義されるエージェントが協調などによって複数で制御するマルチエージェントシステムである．そのほか，従来研究として遺伝的プログラミング[6)]，遺伝的アルゴリズム[7)]，囚人のジレンマ[8)]，強化学習[9)-11)]を導入したマルチエージェントシステムによる交通信号制御に関する研究が行われている．これらの手法は学習により環境に適応する

†1 東京理科大学経営学部
School of Management, Tokyo University of Science
†2 近畿大学理工学部
School of Science and Engineering, Kinki University
†3 金沢大学大学院自然科学研究科
Graduate School of Natural Science and Technology, Kanazawa University

NII-Electronic Library Service

Information Processing Society of Japan

が，解の最適性に重点をおいているものが多い．しかし，エージェントによる学習では学習空間が大きくなるため，多くの学習時間が必要である[12),13)]．そのため，交通信号制御のような実用的な動的環境に導入するためには解の最適性ばかりでなく，学習時間の短縮も重要な課題である．

強化学習の高速化に関する研究は，「学習の収束に要する経験数の減少」と「学習の収束に要する計算時間の短縮」の立場がある[14)]．前者は実機に応用するために少ない経験数で効率的に学習することを目的とし，後者はシミュレーション上の学習における計算時間の短縮を目的とする．前者の立場の学習手法として同一状態の複数のルールを学習する自己強化学習[11)]が提案されているが，状態数が大量の場合，特に学習初期段階では同一状態のルール（学習されたルール）が選択される確率が低いため，学習効果が十分に得られない．また，後者の立場の学習手法として複数のエージェント間で情報を共有する並列計算による手法[14)-17)]である並列強化学習が提案されているが，実機に適用する場合，各エージェントの機能，置かれる環境の差異により，情報の共有が困難な場合がある．これは機能などが類似する他のエージェントを見聞することによって学習する代理強化学習に似ている．このような複数エージェントの情報共有による学習には実機に適用する場合に問題点がある．また，複数のエージェントが存在するマルチエージェント環境でも，他のエージェントの協力が期待できず，単一エージェントごとに学習しなければならないこともある．この場合，他のエージェントの状態は環境の一部として考慮しなければならないため，学習空間は膨大となり，さらに各エージェントが試行錯誤によるランダムな行動をすることは，環境の変化を意味し，また組織的な制御を行う場合には形成し始めた組織を破壊することがあるため，できる限り早く学習によって安定した行動をとるためにも学習の高速化が必要である．

本論文では経験数の減少を目的とした学習手法として，複数エージェント間で行う代理強化学習を単独のエージェント内部で行う学習手法，つまり，類似状態のルールも同時に学習する自己代理強化学習を提案する．この自己代理強化学習を交通信号制御に適用し，シミュレーションを用いて従来手法と比較することによってその有効性を検証する．

2. 大量の状態を必要とする環境における学習方法について

本論文で対象とする強化学習は不確実性や強化信号（報酬）の遅れをともなう不明確な情報によって学習するのが特徴であり，試行錯誤によって学習を行う．また，IF 状態 THEN 行動の IF-THEN 形式のルールを用い，各状態に複数の行動を組み合わせることでルールが作成される．そのため，特に状態数（ルール数）が大量の場合，学習に多くの時間を必要とし，実機に導入するためには学習時間の短縮が要求される．そこで，学習の収束に要する経験数の減少を目的として，学習するルール数に着目し，従来手法（強化学習，自己強化学習，代理強化学習（並列強化学習））と提案手法（自己代理強化学習）について述べる．

2.1 強化学習

強化学習（Reinforcement learning）とは実行したルールに対する環境からの強化信号（Reinforcement）を手がかりに実行したルールを学習する手法である．つまり 1 回のルール実行（経験）で 1 つのルールを学習するため，ルール数が大量の場合，学習過程ではランダム選択による制御が行われる可能性が高く，学習時間も重要である実機に応用する場合には問題となる．

2.2 自己強化学習

自己強化学習（Self reinforcement learning）とは環境からの強化信号の成否を自ら評価し，自ら作成した強化信号である自己強化信号（Self reinforcement）により学習する手法である[18)]．強化学習では環境からの強化信号を刺激のように受け取るため，強化信号は環境に制御されている．しかし，エージェント内部のルール間の関係が明らかであれば，環境から得られた強化信号を評価し，自己強化信号を作成することで関係するルールに強化信号を与えることが可能である．つまり，1 回のルール実行で同一状態のすべてのルールを学習する．そのため，自己強化学習はルールではなく状態に関して学習しており，再び同一状態を認識した場合，強化学習とは異なり認識した状態は学習されている（つまり，1 度認識した状態のルールはすべて 1 度学習されている）ため，ランダム選択によって制御せず，適切な制御が行われる可能性が高い．しかし，学習がある程度行われた段階では学習した状態が再び認識されやすく効果があるが，状態数が大量の場合，特に学習初期段階においては様々な状態を認識するため，学習した状態を認識することが少なく，ランダム選択による制御が行われる可能性が高い．実機に応用する場合，学習初期段階のランダム選択による制御の悪影響がその後の制御に大きく関係することもあるため，学習初期段階であってもランダム選択による制御は抑えるべきであり，対策が必要である．

NII-Electronic Library Service

Information Processing Society of Japan

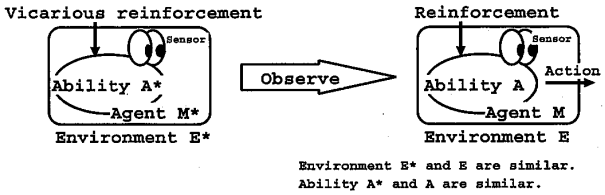


図 1 代理強化学習
Fig. 1 Vicarious reinforcement learning.

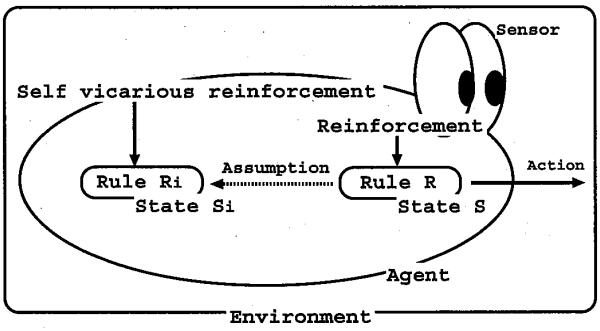


図 2 自己代理強化学習
Fig. 2 Self vicarious reinforcement learning.

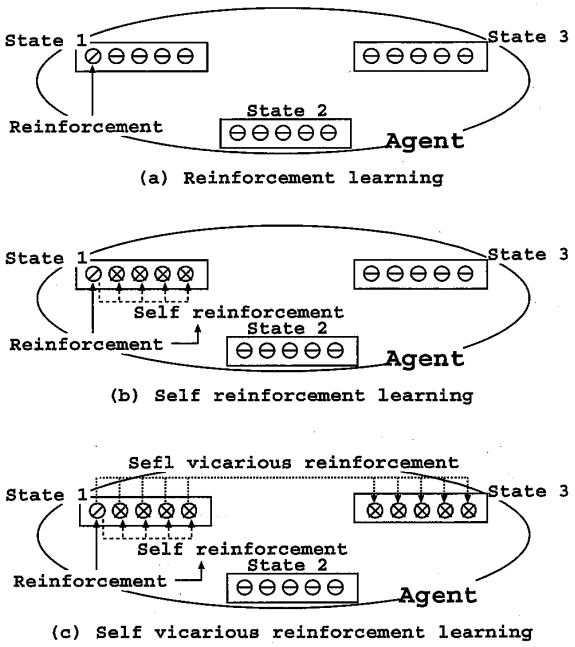


図 3 学習手法の比較
Fig. 3 Comparison between the learning methods.

2.3 代理強化学習（並列強化学習）

代理強化学習（Vicarious reinforcement learning）とは**図 1** に示すように，エージェント M（能力（Ability）：A，置かれている環境（Environment）：E）が受けた強化をエージェント M と類似する状態のエージェント M^*（Ability：A^*，Environment：E^*）が見聞し，エージェント M^* が自分の立場に置き換えて評価することで代理強化信号（Vicarious reinforcement）を作成して学習する手法である[18]．この手法は学習結果を共有する並列強化学習と同様の考え方であり，学習時間の短縮が期待できる．しかし，実機に応用する場合には類似する能力と環境が必要であるという問題点がある．まず，能力に関する条件では，システム設計者が同様の能力となるようにエージェントを作成したとしても，特に実機におけるハードウェア的な面では容易ではなく，コストもかかる．また，環境に関する条件では，環境のすべてを観測することは困難であるため，認識している状態では同様であっても，実際の環境は異なる可能性がある．つまり，代理強化学習は実機に応用するためには厳しい条件を持っており，改善が必要である．

2.4 自己代理強化学習

本論文では 2.2，2.3 節で述べた各学習手法を応用した以下の自己代理強化学習を提案する．

自己代理強化学習

自己代理強化学習とは**図 2** に示すように，エージェント内の状態（State）S が受けた強化を状態 S と類似する状態 S_i が受けた強化と仮定して状態 S_i を評価し，自己代理強化（Self vicarious reinforcement）を作成して状態 S_i を学習する手法である．

図 1 では Agent に関して代理強化学習が行われるのに対し，図 2 に示す自己代理強化学習はエージェント内の State に関して代理強化学習が行われる．つまり，自己代理強化学習は図 1 の Ability を図 2 の Rule とした場合となり，同一の行動部となるルールを学習対象とすれば，代理強化学習の能力に関する条件を満たし，各状態はエージェント内に存在するため，環境に関する条件も満たす．さらに，自己代理強化学習は状態を学習することから 1 つのルール実行において，自己代理強化学習は 2.2 節の自己強化学習後に行う．よって，自己代理強化学習は，1 つのルール実行において，複数の状態の全ルールについて学習する．

図 3 に学習するルール数に着目した場合の強化学習，自己強化学習，自己代理強化学習の様子を示す．図 3 では○は各状態におけるルールを意味する（学習状態により，学習しない場合は ⊖，ルールを実行し，学習する場合は ⊘，ルールは実行しないが学習する場合は ⊗ で表す）．強化学習では，図 3 (a) に示すように 1 つのルール（State1 の ⊘）実行において，その実行したルールに強化信号が与えられ学習する．そのため，同一状態の学習でも，多くの学習回数（図 3 (a) では少なくとも 5 回は State1 について学習しなければならない）が必要となる．自己強化学習では，図 3 (b)

NII-Electronic Library Service

Information Processing Society of Japan

に示すように強化学習（State1 の ⊘ の学習）後，さらに自己強化信号を用いて，同一状態の実行していない複数のルール（State1 の ⊗）を学習する．つまり，1 つのルールではなく 1 つの状態に関して学習する．自己代理強化学習では，図 3 (c) に示すように，自己強化学習（State1 の ⊘，⊗ の学習）後，さらに各ルールの（自己）強化信号に関する自己代理強化信号を用いて，複数の状態（State3）を学習する．

以上より，自己強化学習における学習初期段階でランダム選択による制御が行われる可能性が高いという問題点に対して，自己代理強化学習は 1 つのルール実行において複数の状態を学習することで学習初期段階において学習されている状態を多く存在させることが可能となり，ランダム選択による制御が行われる確率が低下し，学習時間の短縮が期待できる．

上述の自己代理強化学習を適用する際は，類似する状態の設定，自己代理強化信号の作成方法が重要となるが，適用しようとするシステムに関する先行研究が行われていれば，その知識を利用することで設定が可能であると考えられる．そこで，3 章より，交通信号制御に導入した場合について述べる．

3. 自己代理強化学習による交通信号制御

本論文では自己代理強化学習の適用として交通信号制御において車両停止回数を削減する問題を用いる．停止回数を削減するためには隣接交差点との関係が重要であり，隣接交差点の状態を考慮しなければならず，状態数が大量となり，学習に多くの時間が必要である．そのため，学習時間の短縮を目的とする自己代理強化学習の適用は有用である．本論文における信号機エージェント（Traffic Signal Agent）はセンサ（Sensor）によって車両を観測し，推測器（Guess module）では停止車両数を推測する．次に状態認識器（Recognizing module）では停止車両数などから環境を認識し行動の候補となるルール集合を作成し，行動選択器（Selection module）によって行動（青時間：通行権を与える時間）を決め，実行する．最後に学習器（Learning module）によってルールの重みを変更することで学習を行う（図 4）．以下に各構成要素について説明する．

3.1 セ ン サ

SensorOut によって自交差点を通過する車両を観測し，SensorIn によって隣接交差点を通過後，自交差点に進んでくる車両を観測する（図 4 参照）．また，センサは各方向の道路の 2 カ所に設置し，道路ごとに観測を行う（たとえば，碁盤目状の道路形状の場合，4 方向の 8 カ所で車両を観測する）．

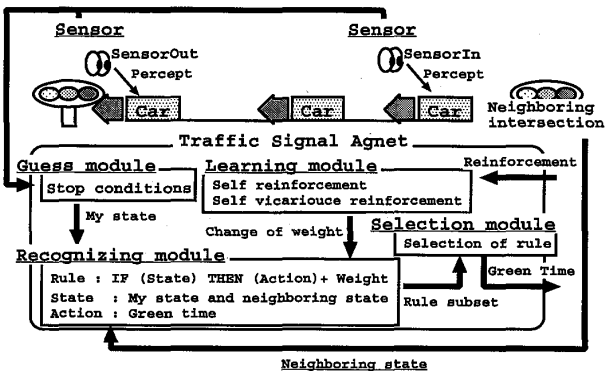


図 4 自己代理強化学習による交通信号機エージェントモデル
Fig. 4 Model of traffic signal agent with self vicarious reinforcement learning.

3.2 推 測 器

車両停止回数を削減するためには，停止車両数（stop conditions）を直接観測することが望まれるが，実際には困難である．そこで，センサ情報から，道路の方向 d，実行した青時間 g 秒後の停止車両数 $mw_d(g)$ を推測する（推測方法については付録参照）．

3.3 状態認識器

IF（状態）THEN（行動）＋（重み）のルールの状態，行動，重みを以下のように設定する．

・状態：自交差点の状態と隣接交差点の状態によって以下のパラメータの値の組合せによって決定される（d：隣接交差点の方向）．

mr： 自交差点の通行権を与えている方向．
mw_d：方向 d の道路の推測停止車両数（$mw_d(0)$）．
nr_d： 方向 d の隣接交差点が通行権を与えている方向．
ng_d： 方向 d の隣接交差点の残り青時間．

隣接交差点の状態（nr_d，ng_d）を用いることにより，隣接交差点を考慮した学習が可能となる．しかし，隣接交差点数を N_n とすると，ルールの状態数は $mr\prod_{d=1}^{N_n}(mw_d*nr_d*ng_d)$ と大量となり，学習が困難である．この対策として学習器を分割する Modular-Q-Learning[19] が有効であり，本論文では方向 d ごとに学習器を分割する．よって方向 d の学習器の状態数は $mr(mw_d*nr_d*ng_d)$ となり，効率的な学習が可能である．

・行動：以下のパラメータを学習する．

g_j：青時間（j はルール番号）．

従来の交通信号制御では 3 つのパラメータ（サ

NII-Electronic Library Service

Information Processing Society of Japan

イクル，スプリット，オフセット）により制御する．しかし，複数のパラメータを学習することは困難であり，学習速度低下の原因となる．そこで，行動として青時間を用いる．

・重み：交差点を停止せずに通過した車両台数により重みを更新する．

3.4 行動選択器

本論文ではルールの重みとして停止せずに交差点を通過した車両台数を用い，また学習器を方向 d ごとに分割するため，各方向 d の学習器の重みの和を用いて，最も重みの重いルールを選択する．つまり，ある状態における方向 d のルール番号 j $(j=1,2,\cdots)$ の重みを W_j^d とすると，$\max\{\sum_d W_1^d, \sum_d W_2^d, \cdots\}$ を満たすルール j の青時間 g_j を実行する．

3.5 学習器

2.4 節で提案した自己代理強化学習を用いる．学習は方向 d ごとに行い，重みの初期値は 0 である．以下，自己強化学習，自己代理強化学習について述べる．

[自己強化学習]

ルール実行中に該当するルールの重みを更新し，ルール実行後は推測により重みを更新する．以下にアルゴリズム SRL4TSC（Self Reinforcement Learning for Traffic Signal Control）を示す．

アルゴリズム SRL4TSC

実行する青時間を g_s 秒とする（s は選択されたルール番号）．

(1) 青時間 g_j 秒のルール（行動部分が g_j 秒になっているルール）ごとに以下の報酬 sr_{dj} を与える．

・通行権を与えている方向の場合

$$sr_{dj} = (tn_d(g_j) - mw_d(0)) * \beta$$

・通行権を与えていない方向の場合

$$sr_{dj} = -mw_d(g_j)$$

ただし，$tn_d(g_j)$ は方向 d における g_j 秒間の SensorOut の観測車両数であり，青時間 g_s 秒実行後は推測観測車両数 $(tn_d(g_s) + mw_d(g_j))$ である（ただし，$mw_d(g_j)$ は付録の停止車両推測式の $sd[i,g]$ の $i * cs$ を省いた式を用いる）．また，β (>1) は定数である．

(2) 初回学習時は報酬をそのまま重みとし，その他の場合は平均により重みを更新する．

本論文では，停止回数削減が目的であるため，停止車両数を負の報酬，SensorOut によって観測される自交差点の通過車両数を正の報酬とした．つまり，通行権を与えている方向では待ち行列（信号機による停止車両数）分の負の報酬から始まるが，待ち行列の車両通過後は正の報酬（停止せずに通過した車両）が増加し，逆に通行権を与えていない方向は車両が自交差点を通過しないため負の報酬が増加する．ここで，通行権が与えられている方向で負の報酬のルールを実行することは，停止していた車両を再び停止させることになる．そのため，通行権を与えていない場合より負の報酬を大きくすべきである．また，通過車両数の影響を大きくする方が良いと考えられる．そこで，通行権を与えている方向の報酬を β 倍している．

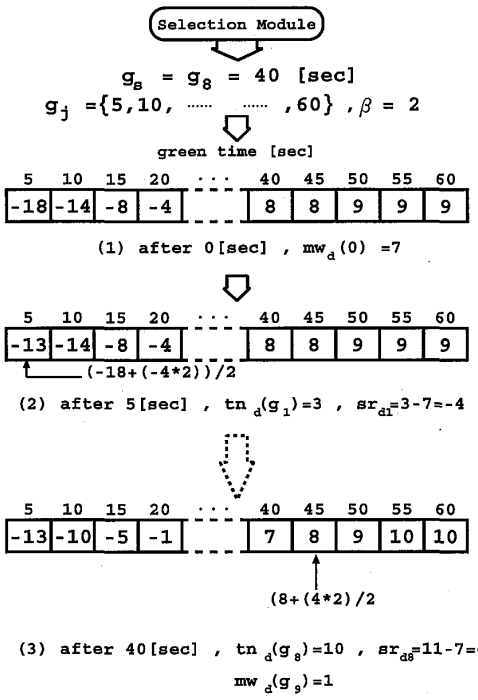


図 5 自己強化学習例

Fig. 5 Example of self-reinforcement learning.

アルゴリズム SRL4TSC の具体例を**図 5** に示す（図 5 では青時間 g_j は 5 秒ごとに設定している）．図 5 では，青時間 g_8 $(=40)$ 秒が選択され，通行権を与えている方向 d の重みがボックス内の数値であるとする．まず，通行権を与える時間を変えたときに $mw_d(0)$ $(=7)$ を計算する（図 5(1)）．青時間 g_1 $(=5)$ 秒実行後，$tn_d(g_1) = 3$ であったとすると報酬は $sr_{d1} = 3 - 7 = -4$ となり，すでにある重み (-18) との平均により重みを更新する（図 5(2)）．これを選択した青時間が終了するまで繰り返す．さらに，青時間実行終了後は，たとえば，$tn_d(g_8) = 10$，$mw_d(g_9) = 1$ ならば，推測観測車両数は $tn_d(g_8) + mw_d(g_9) = 10 + 1 = 11$ となり，報酬は $sr_{d9} = (10+1) - 7 = 4$ となり，前述のルールと同様に重みを更新する（図 5(3)）．これは青時間実行中に学習されていないルールすべて（図 5 では青時間が 45，50，55，60 秒のルール）に適用する．

NII-Electronic Library Service

Information Processing Society of Japan

[自己代理強化学習]

自己強化学習後，学習したルールの状態と類似する状態のルールの重みを更新する．本論文ではルールの状態のパラメータ mw_d のみが異なり，かつ他のパラメータが等しい状態を類似状態とする．以下にアルゴリズム SVRL4TSC（Self Vicarious Reinforcement Learning for Traffic Signal Control）を示す．

アルゴリズム SVRL4TSC

(1) 実行されている状態の推測停止車両数を mw_{de}，自己代理強化学習を行う状態の推測停止車両数を mw_{dv}（$mw_{de} \neq mw_{dv}$）とする．

(2) mw_{de} の状態に対して g_j 秒ごとに作成された自己強化学習による報酬 sr_{dj} を用いて $vr_{dj} = sr_{dj} - (mw_{dv} - mw_{de})$ を mw_{dv} の状態の報酬とし，重みを更新する．

ここで，重みの更新については自己強化学習と同様である．ただし，同一状態の 2 回目以降の学習では，過去に自己強化学習が行われた状態には自己代理強化学習は適用せず，初めて自己強化学習が行われる場合は，自己代理強化学習の結果は考慮せず，学習されていない状態として扱う．

本研究における類似状態の決定は以下の理由による．類似状態の対象とした mw_d は推測停止車両数である（3.3 節参照）．また，自己強化学習における重みの更新でも mw_d が用いられている（3.5 節参照）．そのため，状態認識時の mw_d の違いは，そのまま学習（重みの更新）にも影響を与えることが分かる．つまり，上記の自己代理強化学習の対象となる類似状態は通行権が変わるときの停止車両数である待ち行列（$mw_d(0)$）が異なる状態を対象としている．よって，本論文における学習では待ち行列は初期の負の報酬であり，待ち行列が異なることは初期の負の報酬が異なることを意味し，その負の報酬の違いが $mw_{dv} - mw_{de}$ で表されている．このように自己代理強化学習は，実際にはまだ認識していない状態を学習するが，状態の違いと報酬（重みの更新）に与える影響の関係が明確な状態を類似状態として選ぶことによって複数状態の学習が可能となる．

アルゴリズム SVRL4TSC の具体例を**図 6** に示す．図 6 では mw_{de}（$= 3$）の状態を認識し，この状態に対して得られた報酬 sr_{dj} を用いて，mw_{dv}（$= 9$）の状態に $vr_{dj} = sr_{dj} - (9 - 3)$ の報酬を与える．

4. シミュレーション実験

本論文で提案した自己代理強化学習による交通信号制御の有効性を検証するためにシミュレーション実験を行った．以下に概要を示す．

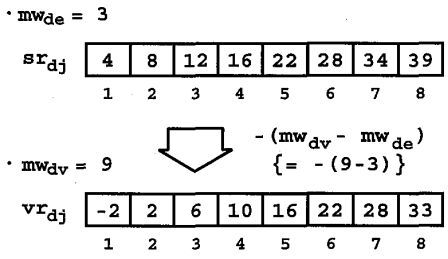


図 6 自己代理強化学習例

Fig. 6 Example of self vicarious reinforcement learning.

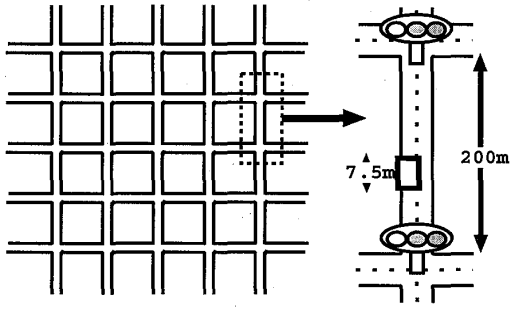


図 7 シミュレーション環境

Fig. 7 Simulated environment.

4.1 シミュレーション概要

道路環境は以下のとおりである（**図 7** 参照）．

道路環境

道路形状	：碁盤目状
交差点数	：5 × 5 交差点
交差点間隔	：200 m
車線	：片側 1 車線（右折レーン容量 5 台）
制限速度	：約 35 km/h
車両長	：7.5 m

シミュレーション環境として実際の道路環境を用いることが望ましいが，交通量の調査などにより困難である．本論文では，自己代理強化学習の有効性を検証するために，シミュレーションのテストベッドとして用いられる碁盤目状の環境とした．また，車両の動作はセルオートマトン法に基づき，不確実性として揺らぎなどを導入し[20]，実験環境の外から流入する車両はポアソン分布に従うものとする[1]．

また，以下の制御方式による比較を行った．

比較制御方式

- INC40：青時間 40 秒，隣接信号機と同期をとらない信号機群による制御．つまり，隣接信号機との関係はランダムである．
- CNC40：青時間 40 秒，隣接信号機と同期をとる信号機群による制御．つまり，全信号機が同時に通行権が同じ方向に変わる．

NII-Electronic Library Service

Information Processing Society of Japan

・ISRL：文献11)により提案されている信号機群による制御．この信号機は待ち時間削減を目的とする自己強化学習を行うが，隣接信号機の状態を考慮していない．
・CSRL：停止回数削減を目的とする自己強化学習（SRL4TSC）を行う信号機群による制御．
・CSVRL：提案手法である停止回数削減を目的とする自己代理強化学習（SVRL4TSC）を行う信号機群による制御．

提案手法に関する状態の分割とパラメータ値は経験的に以下のように設定した．

$$
\begin{aligned}
d &= \{\text{東，西，南，北}\}\\
mr &= \{\text{東西，南北}\}\\
mw_d &= \{mw_d(0) \le 6, 6 < mw_d(0) \le 12,\\
&\quad 12 < mw_d(0) \le 18, 18 < mw_d(0)\}\\
nr_d &= \{\text{東西，南北}\}\\
ng_d &= \{0, (5,10), (15,20), (25,30),\\
&\quad (35,40), (45,50), (55,60)\}\\
g_j &= \{5, 10, 15, 20, 25, 30, 35, 40, 45,\\
&\quad 50, 55, 60\}\\
\beta &= 2
\end{aligned}
$$

上記の環境において制御方式ごとに20回のシミュレーションを行い，その平均により評価を行う．以下にシミュレーション結果を示す．以下のシミュレーション結果の図では縦軸が交差点を通過するまでの車両1台あたりの平均停止回数，横軸がシミュレーション時間である．

4.2 シミュレーション結果1

各交差点での車両直進率90%（右折，左折率各5%），すべての方向からの環境への車両流入率が1台/6秒の条件によるシミュレーションを行った．この環境では，各交差点の通行権の方向により交通流に偏りがあり，隣接交差点の影響が大きく，その影響を考慮した学習が必要である．つまり，学習に多くの時間が必要となる．

シミュレーション結果を**図8**に示す．図8より，提案手法であるCSVRLが最も平均停止回数が少なく，続いてCSRL，CNC40，INC40，ISRLとなることが分かる．

INC40，CNC40の比較から，この2つの制御方式の違いは隣接信号機との同期の有無であり，CNC40の方が平均停止回数が約0.1[回/台]ほど良い結果であることから，本シミュレーション環境では隣接交差点との適切な関係が必要であることが分かる．また，学習システムであっても隣接交差点の状態を考慮していない

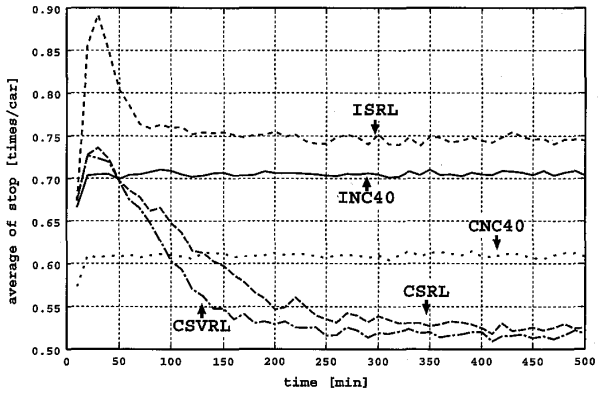


図8 シミュレーション結果1
Fig. 8 Results of the simulation 1.

表1 目標停止回数によるCSRLとCSVRLの比較
Table 1 Comparison between CSRL and CSVRL in stop conditions with goal.

goal[times/car]		0.70	0.65	0.60	0.55
goal	CSRL	50	100	150	230
time[min]	CSVRL	50	80	110	140

ISRLが平均停止回数が約0.75[回/台]と最も悪い結果であることからも，隣接交差点の考慮は重要であることが分かる．次にCSRLとCSVRLの比較を行う．この2つの制御方式はどちらも隣接交差点の状態を考慮しているが，学習手法が異なる．両制御方式を比較すると，50分まではほとんど違いが見られないが，時間が経過するにつれて差が大きくなることが分かる．ここでは段階的に目標停止回数（goal）を設定し，その目標の達成時間（goal time）によって比較を行う．結果を**表1**に示す．表1から分かるように，最初は差がないが，目標停止回数が小さくなるにつれて，目標達成時間の差は大きくなり，最終的には90分の差が生じ，自己代理強化学習は自己強化学習の$140/230 \simeq 52\%$に学習時間が短縮されている．また，学習結果もCSVRLの方がCSRLより約0.015[回/台]ではあるが良いことが分かる．これは，学習初期段階（本シミュレーションでは50〜100分）における制御の適切さが影響していると考えられ，交通信号制御のように渋滞などの影響が一時的なものではない場合には自己代理強化学習は有効であることが分かる．

4.3 シミュレーション結果2

南北方向からの流入率が1台/5秒，東西方向からの流入率が1台/50秒，交差点での車両直進率90%の条件によるシミュレーションを行った．この環境では，各方向によって隣接交差点の影響が異なり，各方向に応じた学習が必要である．

シミュレーション結果を**図9**に示す．この環境では

NII-Electronic Library Service

Information Processing Society of Japan

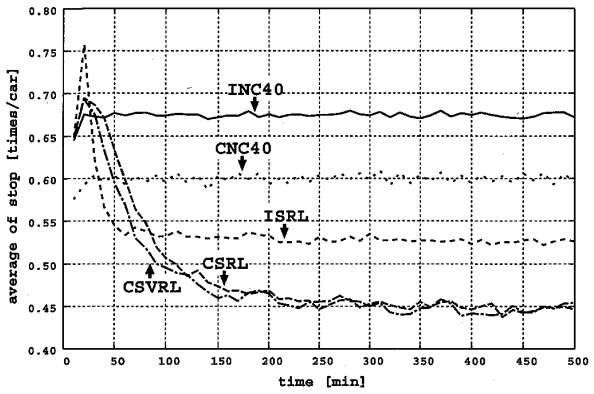


図 9 シミュレーション結果 2
Fig. 9 Results of the simulation 2.

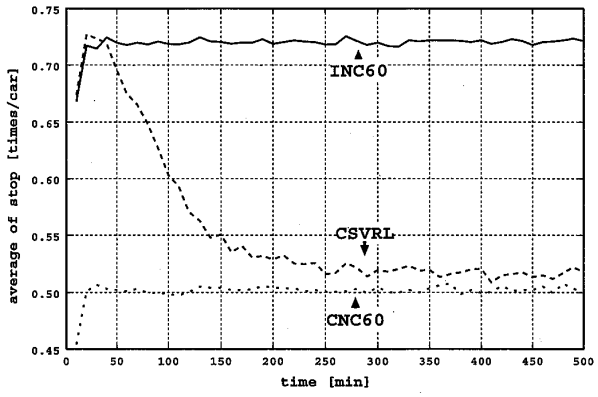


図 10 解の最適性の検証 1
Fig. 10 Verification of optimality 1.

CSRL，CSVRL が最も平均停止回数が少ないことが分かる．また，隣接交差点を考慮していない ISRL が CNC40 よりも平均停止回数が少ないことが分かる．よって上記の条件では，隣接交差点を考慮した制御よりも交通量の多い方向の通行権を長く与えた方が全体としての平均停止回数を減少させることができる．また，自己代理強化学習の効果が顕著に得られないのは，流入率の小さい方向の学習が迅速に行われ，4.2 節の環境に比べ多くの状態を認識しないためであると考えられる．

4.4 シミュレーション結果 1，2 に関する考察

本研究では選択する青時間によって学習回数は異なるが，シミュレーション結果 1 の条件では平均すると CSRL が 520.9 回，CSVRL が 513.2 回の学習回数であり，有意差はない．しかし，1 回に学習する状態（ルール）の数が異なるため，4.2，4.3 節より静的な環境では自己代理強化学習により学習時間が短縮できることが分かった．以下に解の最適性，組織的制御，ランダム選択率に関して考察を行う．

4.4.1 解の最適性

本論文では解の最適性については考慮していないが，良い学習結果が得られることが望ましい．4.2 節では提案手法による学習の結果，90%は青時間 55 秒以上のルールを選択し，制御を行っていた．そこで，解の最適性を検証するために，INC40，CNC40 の制御の青時間を 60 秒とした INC60，CNC60 についてシミュレーションを行い，比較した．結果を**図 10** に示す．本研究におけるシミュレーションでは交差点での車両直進率が 90%であるため，隣接信号機とできる限り長い時間同じ方向に通行権を与えている方が車両を停止させずに通過させることができる．よって，CNC60 は同期をとるため，最適解の 1 つであると考えられる．図 10 から，提案手法である CSVRL は学習初期段階では INC60 とほぼ同じであるが，学習が行われるにつれて学習結果は最適解である CNC60 に接近しており，十分な学習結果が得られている．さらに，4.3 節の流入率が不均質の学習結果の最適性を検証するために，CSVRL の学習結果（南北方向青時間 60 秒，東西方向青時間 15 秒）を用いた信号機 CNC6015 についてシミュレーションを行い，比較した．結果を**図 11** に示す．図 11 から CSVRL の方が平均停止回数が少ないことが分かる．つまり，より複雑な制御が行われており，高い有効性があると考えられる．

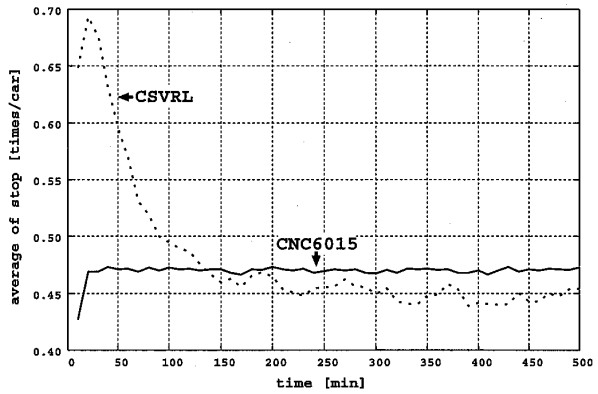


図 11 解の最適性の検証 2
Fig. 11 Verification of optimality 2.

4.4.2 組織的制御

4.4.1 項から，本提案手法による交通信号制御では学習の結果，隣接交差点を考慮した組織的（系統）制御が行われていると考えられる．そこで，隣接交差点との関係を検証するために 100 分ごと（0〜100 分，101〜200 分，…）のオフセット（隣接信号機との青信号開始時間のずれ［秒］）の平均を**表 2** に示す（つまり，CNC40，CNC60 はオフセット 0 秒となる）．表 2 から学習が進むにつれてオフセットが小さくなっており，隣接交差点と同期をとるように制御が行われていることが分かる．また，CSVRL の方が収束が早いことが分かる．つまり，本研究では明示的に同期をとることを行っていないが，学習によって CNC60 に近づくように同期するように学習していると考えられる．本研

NII-Electronic Library Service

Information Processing Society of Japan

表 2 100 分ごとのオフセットの平均 [秒]
Table 2 Average of the offset.

simulation time	100	200	300	400	500
CSRL	23.3	15.0	12.3	11.6	9.3
CSVRL	21.0	8.1	6.2	4.9	5.8

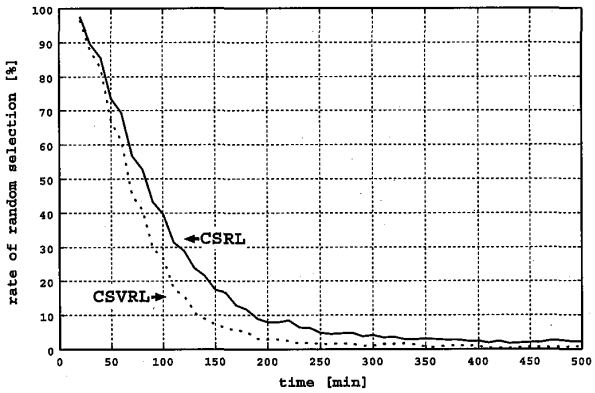


図 12 ランダム選択率による CSRL と CSVRL の比較
Fig. 12 Comparison between CSRL and CSVRL in random selection.

表 3 目標ランダム選択率による CSRL と CSVRL の比較
Table 3 Comparison between CSRL and CSVRL in random selection with goal.

goal[%]		70	50	30	10
goal	CSRL	60	90	120	190
time[min]	CSVRL	50	70	90	140

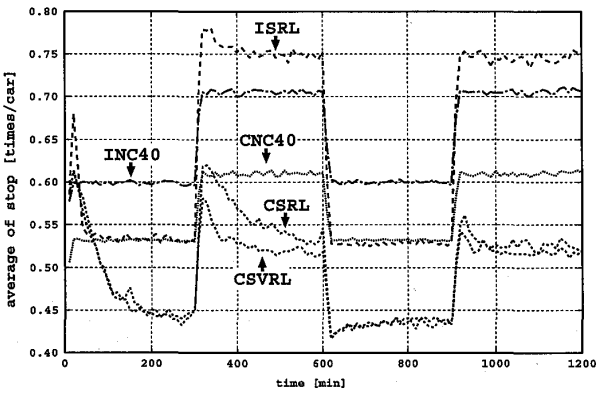


図 13 シミュレーション結果 3-1
Fig. 13 Results of the simulation 3-1.

究では，このように形成された制御を組織的な制御と呼ぶ．

4.4.3 ランダム選択率

次に自己強化学習と自己代理強化学習についてランダム選択率（ルールをランダムに選択する確率）を比較する．ランダム選択率は選択した全ルール数のうちのランダムに選択したルールが占める割合である．4.2 節の場合の比較結果を**図 12** に示す．全体を通して，CSVRL は CSRL よりランダム選択率が小さいことが分かる．ここでも，停止回数と同様に段階的に目標ランダム選択率を設定し，その達成時間を比較する．結果を**表 3** に示す．表 3 から，学習が行われるにつれて達成時間の差が大きくなるが，表 1 に比べると小さいことが分かる．これは CSRL，CSVRL は隣接交差点を考慮しているため，最終的には隣接交差点に応じた組織的な制御を行うが，少しずつ組織を形成していくと考えられる．学習初期段階では，ランダム選択率が高いため組織的な制御は形成されない．しかし，学習中期段階（本シミュレーションでは 100〜150 分）では，組織的な制御が行われ始めるため，ランダム選択が行われると，形成され始めた組織的な制御が破壊されてしまうと考えられる．破壊された組織的な制御はすぐにもとに戻すことはできず，その結果，ランダム選択率の差以上の影響が制御（学習速度）に影響していると考えられる．

4.5 シミュレーション結果 3

4.2〜4.4 節より静的環境において提案手法である自己代理強化学習は，学習時間短縮に有効であり，また組織を形成するような交通信号制御に有効であると考えられる．しかし，実環境は静的ではなく動的であることが多いことから，環境を変化させた場合のシミュレーションを行う．動的環境として交通量を 300 分ごとに {少→多→少→多} の状態に変化させた場合，流入率の高い方向を 300 分ごとに {南北→東西→南北→東西} の状態に変化させた場合のシミュレーションを行った．

交通量を変化させた場合のシミュレーション結果を**図 13** に示す．図 13 から CSVRL はすべての時間帯において最も少ない平均停止回数であり，良い結果となっていることが分かる．特に最初の交通量の変化（300 分時）に対して，CSVRL が平均停止回数の増加が少なく，迅速に環境に適応していることが分かる．この環境では直進率が等しく，交通量のみが変化するため信号機群の組織的な制御を大きく変更する必要はない．このとき，交通量が変わることから，信号機エージェントは今までに経験していない環境に置かれ，新しい状態を認識する．しかし，CSVRL は自己代理強化学習によって交通量と関係する待ち行列の類似状態に対して学習が行われているため，環境に迅速に適応している．このときの環境変化時における CSRL，CSVRL のそれぞれのランダム選択率の変化は 3% → 25%，1% → 19% であり，CSVRL の方が形成した組織を壊さずに制御を行っていることを確認した．

流入率を変化させた場合のシミュレーション結果を

NII-Electronic Library Service

Information Processing Society of Japan

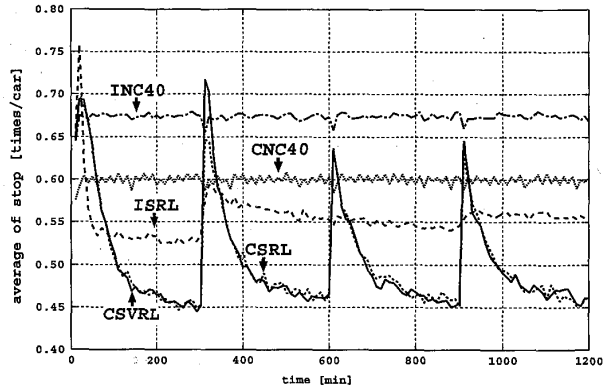


図 14 シミュレーション結果 3-2
Fig. 14 Results of the simulation 3-2.

図 14 に示す．図 14 から CSRL，CSVRL が適していることが分かるが，環境変化ごとに平均停止回数がいったん増加していることが分かる．この環境では，南北方向の流入率が多い場合は南北方向，東西方向の流入率が多い場合は東西方向に重点をおいた組織的な制御が必要であるため，環境変化ごとに組織の形状を大きく変化させなければならない．そのため，組織を再形成する過程で平均停止回数が増加してしまい，学習に時間を要している．また，1 回目の環境変化時に CSVRL の方が CSRL より平均停止回数が多くなっていることが分かる．この環境で起こる環境変化は信号機エージェントが認識していない（隣接交差点からの車両流入に関する）状態であるため，信号機エージェントが同様の状態を認識しても，実際の環境は異なる．そのため，自己代理強化学習によって待ち行列の類似状態について学習しているが，実際は異なる環境であるため逆効果となっている．しかし，2 回目以降では自己強化学習によって得られる学習結果が増加するため，自己代理強化学習によって得られた学習結果を使うことが少なくなることから，CSRL とほぼ同じ結果になっており，逆効果は一時的なものであると考えられる．

5. 関連研究

本研究のように類似状態が定義できる場合，状態空間の縮約による学習の高速化が考えられる．しかし，単純に状態空間を縮約した場合には学習の高速化は期待できるが，性能の低下が起こるであろうことが予想される．たとえば，類似状態であってもその状態に適した行動は異なる可能性があることが考えられる*1．

*1 これに関しては文献 21) において考察されており，知覚制限（状態数削減）条件で獲得した評価値を完全知覚での評価関数の初期値として用いた場合には学習の高速化は達成できるが，学習性能が低下していることが示されている．

文献 21) ではマルチエージェント強化学習における状態数の削減（知覚情報の粗視化）により学習の高速化を行っている．具体的には知覚制限学習器と完全知覚学習器により並列学習を行い，学習の初期段階では知覚制限学習器によって学習を高速化し，適切なタイミングで完全知覚学習器に切り替えることで学習性能も維持する手法である．文献 21) では切替えのタイミングに関して様々な考察が行われているが，切替え時に性能が低下するという問題点がある．これは交通信号制御問題のように少しの失敗が後になって大きな影響を及ぼす可能性がある場合には致命的であり，改善が必要である．これに対して，本研究における自己代理強化学習では学習途中で性能が低下することがなく，また切替えのタイミングを考慮する必要がない．しかし，類似状態の決定を考慮する必要があり，類似状態の決定に関しては適用問題に依存するところが大きいが，エージェント設計時（認識する状態の決定や報酬の与え方など）において，状態や報酬の間に関連性が生じること（本研究における mw_d と報酬の関係）があり，これを利用することで，その他の問題にも適用できると考えられる．また，初めての環境変化時に性能が低下する場合があり，この問題に関しては自己代理強化学習の結果を利用するかどうかを判断するなどの改善が必要である．

6. まとめ

本論文では，実環境における学習時間短縮手法として自己代理強化学習を提案し，交通信号制御へ応用した．交通信号制御の目的を停止回数削減とする場合，隣接交差点の状態を考慮するため，大量の状態について学習しなければならず，多くの学習時間が必要となる．シミュレーションの結果，自己代理強化学習の導入により，従来手法である自己強化学習の約半分の時間で学習が可能であり，さらに学習結果も改善されることを示した．また，動的環境に関するシミュレーションを行い，その特性を検証した．

今後は，より実際的な環境においてシミュレーションを行い，どのような環境にどのような制御が適しているか検証する予定である．

参考文献

1) 猪瀬 博，浜田 喬：道路交通管制，産業図書 (1972).
2) 岡本博之：道路交通の管理と運用，技術書院 (1987).
3) 佐佐木綱，飯田恭敬：交通工学，オーム社

NII-Electronic Library Service

Information Processing Society of Japan

(1992).
4) IBM：CM "HELP DESK2" シリーズ「交通渋滞」編.
5) 国土交通省道路局：ITS ホームページ.
6) 余田精一，渥美雅保：遺伝的プログラミングに基づく交通信号制御プログラムの協調学習，人工知能学会全国大会，pp.175–178 (1996).
7) 山本直史，森下 信：セルラオートマトンによる交通流のモデル化とその制御，日本機械学会機械力学・計測制御講演論文集，Vol.B, pp.265–268 (1998).
8) 堂前卓也，遠藤聡志，山田孝治：マルチエージェントシステムに関する基礎研究—交通信号制御へのアプローチ，映像情報メディア学会技術報告，Vol.22, No.33, pp.67–72 (1998).
9) 三上貞芳，高橋真紀，和田充雄：マルチエージェント強化学習とその産業応用，電気学会産業システム情報化研究会資料 IIS-98-38, pp.33–38 (1998).
10) 吉田 功，山村雅幸：交通システムにおける適応的信号制御，第 26 回知能システムシンポジウム，pp.157–162 (1999).
11) 参沢匡将，木村春彦，広瀬貞樹，大里延康：強化学習型マルチエージェントによる交通信号制御，電子情報通信学会，Vol.J83-D-I, No.5, pp.478–486 (2000).
12) 荒井幸代，宮崎和光，小林重信：マルチエージェント強化学習の方法論—Q-Learning と Profit Sharing による接近，人工知能学会誌，Vol.13, No.4, pp.609–617 (1998).
13) 三上貞芳：強化学習のマルチエージェント系への応用，人工知能学会誌，Vol.12, No.6, pp.845–849 (1997).
14) 森紘一郎，山名早人：強化学習並列化による学習の高速化，情報処理学会研究報告 (ICS)，Vol.2004, No.29, pp.89–94 (2004).
15) Kretchmar, R.M.: Parallel Reinforcement Learning, *The 6th World Conf. on Systemics, Cybernetics, and Informatics* (2002).
16) Antonova, D.: Parallel Reinforcement Learning—Extending the Concept to Continuous Multi-State Tasks, thesis, Denison University (2003).
17) Kaya, M. and Arslan, A.: Parallel and distributed multi-agent reinforcement learning, *Proc. 8th International Conference on Parallel and Distributed Systems* (*ICPADS 2001*), pp.437–441 (2001).
18) 梅本堯夫，大山 正：心理学への招待，サイエンス社 (1992).
19) Ono, N. and Fukumoto, K.: A Modular Approach to Multi-agent Reinforcement Learning, *Distributed Artificial Intelligence Meets Machine Learning*, pp.25–39 (1997).
20) 加藤恭義：セルオートマトン法による道路交通シミュレーション，人工知能学会誌，Vol.15, No.2, pp.242–250 (2000).
21) 伊藤 昭，金渕 満：知覚情報の粗視化によるマルチエージェント強化学習の高速化—ハンターゲームを例に，電子情報通信学会論文誌，Vol.J84-D-I, No.3, pp.285–293 (2001).

付 録

A.1 停止車両数の推測アルゴリズム

本論文では以下の式により，方向 d，実行青時間 g 秒後の停止車両数 $mw_d(g)$ を推測する（図 **15** (a) 参照）.

$mw_d(g) = i_{max} + 1$

($i_{max} : sd[i,g] \leq 0$ となる i の最大値)

$sd[i,g] = rs - i * cs - (pt(g) - R[i]) * (\alpha * ls)$

g : 実行した青時間 [s]

d : 推測しようとする道路の方向

i : 対象とする道路の先頭車両を 0 とした場合の車両番号 ($i = 0, 1, \cdots$)

rs : 道路長（交差点間距離）[m]

cs : 車両長（停止車間距離を含む）[m]

ls : 制限速度 [m/s]

$pt(g)$: 基準となる時間から通行権を変えたときまでの経過時間 $+g$[s]

$R[i]$: 車両番号 i の隣接交差点通過時間

$\alpha(< 1)$: 定数

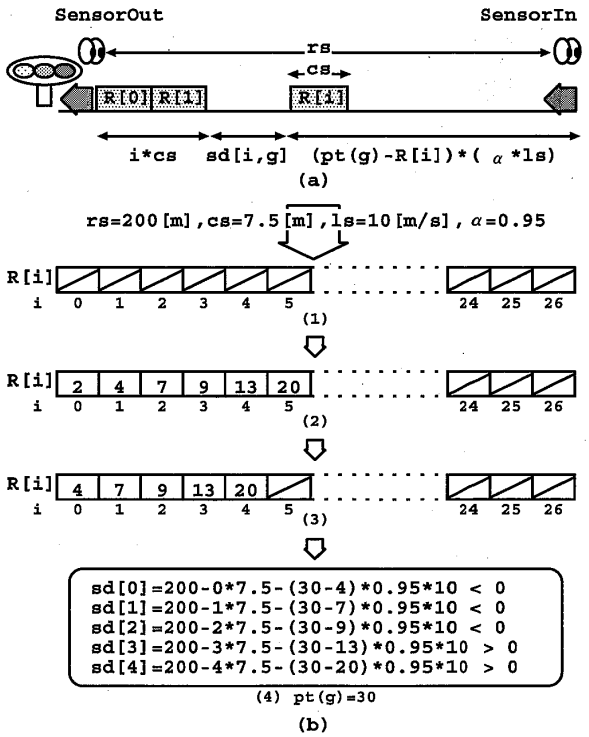


図 **15** 停止車両の推測
Fig. 15 A guess of stop condition.

NII-Electronic Library Service

Information Processing Society of Japan

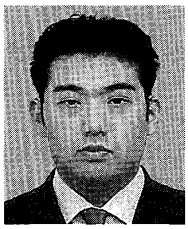
上式では，$pt(g) - R[i]$ は車両番号 i の車両が隣接交差点を通過してからの経過時間であるため，$(pt(g) - R[i]) * (\alpha * ls)$ は隣接交差点からの走行距離である．ここで，定数 α は車両が必ずしも制限速度で隣接交差点から流入するとは限らないことから用いている．また，$i*cs$ は車両番号 i の車両が通過しようとする交差点で停止する場合の交差点からの距離である．よって，図 15 (a) から分かるように，$sd[i,g]$ は車両番号 i の車両が停止するまでに可能な走行距離を示しており，$sd[i,g] \leq 0$ ならばすでに車両番号 i の車両は停止していると考えられる．つまり，$sd[i,g] \leq 0$ を満たす i の最大値 $(i_{max}) + 1$ が交差点間の推測停止車両数となる．

具体例を図 15 (b) に示す．図 15 (b)(1) に示すように車両通過時間を記録するための配列を用意し，車両が隣接交差点から流入（SensorIn が車両を観測）するごとに通過時間を記録していく（図 15 (b)(2)）．また，車両が自交差点を通過（SensorOut が車両を観測）した場合は，$R[i]$ を更新する（図 15 (b)(3)）．たとえば，$pt(g) = 30$ とすると，図 15 (b)(4) に示すように $sd[i]$ が計算されるため，$sd[i] < 0$ を満たす i_{max} は 2 であることから推定停止車両数は 3 台となる．また，本論文におけるシミュレーションでは経験的に $\alpha = 0.95$ とした．

(平成 18 年 11 月 10 日受付)
(平成 19 年 10 月 2 日採録)

参沢 匡将（正会員）

平成 11 年金沢大学工学部電気・情報工学科卒業．平成 16 年同大学大学院自然科学研究科博士後期課程数理情報科学専攻修了．博士（工学）．同年金沢工業大学高度材料科学研究開発センター特別研究員．平成 17 年東京理科大学経営学部助手．平成 19 年同助教．現在に至る．マルチエージェント，強化学習，ペットロボット，人工市場に関する研究に従事．電子情報通信学会，人工知能学会，電気学会各会員．



阿部 孝司（正会員）

平成 8 年工学院大学情報工学科卒業．平成 10 年同大学大学院修士課程修了．平成 13 年金沢大学大学院自然科学研究科博士後期課程修了．博士（工学）．同年同大学工学部付属電磁場制御実験施設講師．平成 14 年英国ノーザンブリア大学画像データ研究所研究員．平成 15 年金沢工業大学情報工学科講師．平成 18 年近畿大学理工学部情報学科講師．現在に至る．マルチメディア情報検索，医用画像処理に関する研究に従事．IEEE，電子情報通信学会，電気学会各会員．

下川 哲矢

平成 12 年東京大学大学院経済学研究科博士課程修了．経済学博士．平成 9 年日本学術振興会特別研究員 DC．平成 11 年東京大学日本経済国際共同研究センター研究員．平成 12 年日本学術振興会特別研究員 PD．平成 15 年東京理科大学経営学部専任講師．現在に至る．ファイナンス理論研究，ファイナンス統計学研究，金融資産価格予測システム構築に従事．

木村 春彦（正会員）

昭和 49 年東京電機大学工学部応用理化学科卒業．昭和 54 年東北大学大学院情報工学専攻博士課程修了．工学博士．同年富士通（株）入社．昭和 55 年金沢女子短期大学講師，昭和 56 年同短期大学助教授，昭和 59 年金沢大学経済学部助教授，平成 4 年同大学工学部電気・情報工学科助教授，平成 6 年同学科教授．現在に至る．その間，最適コード変換，プロダクションシステムの高速化の研究に従事．電子情報通信学会，人工知能学会各会員．

NII-Electronic Library Service