分散深層強化学習を用いたモバイルデータオフローディング手法の提案

望月大輔[1] 安孫子悠[2] 峰野博史[3]

概要： IoT（Internet of Things）の普及に伴い，モバイルデータ通信の需要は今後も増え続けると予想される．通信キャリアは Wi-Fi スポットを設置しモバイルデータ通信網の負荷を分散するモバイルデータオフローディングに取り組んでいる．帯域利用効率を最大化する方法として，遅延を許容する即時性を求めないデータに着目し，送信レートを制御することで帯域利用効率を最大化することを目的とした Mobile Data Offloading Protocol（MDOP）が提案されている．しかし，MDOP の時間的オフローディングは送信レート制御に定式化された数理モデルを用いており，多種多様な通信インフラの状況に応じて常に帯域利用効率を最大化することは困難である．本研究では，多種多様な通信インフラの状況において常に帯域利用効率を最大化するため，分散深層強化学習を用いたモバイルデータオフローディング手法を提案する．基礎検討として，MDOP の時間的オフローディングに焦点を当て強化学習の適用可能性について評価を行った．評価の結果，分散深層強化学習を用いた送信レート制御モデルが，時間的局所性を解消するような送信レート制御を行うことが可能となり，単一の送信レート制御モデルに比べ帯域利用効率を 6%向上させられることを確認した．

Mobile Data Offloading Protocol using distributed deep reinforcement learning

DAISUKE MOCHIZUKI[1] YU ABIKO[2] HIROSHI MINENO[3]

1. はじめに

Internet of Things（IoT)の普及や携帯端末の性能向上によるコンテンツの多様化に伴い，モバイルデータ通信需要は増加傾向にある．モバイルデータ通信需要は，2021 年には約 7 倍になると予想されている[1]．特に，今後急激に増加すると予想されるデバイスとして，人が操作することなく個別に稼働する機器同士が自律的に制御を行う Machine-to-Machine（M2M）端末が挙げられ，膨大なモバイルデータトラフィックに対して，通信インフラの帯域を効率的に使用することが重要となる．通信キャリアは通信インフラの負荷を分散するため，Wi-Fi スポットを設置しモバイルデータオフローディングに取り組んでいる．今日のモバイルデータ通信の特徴として，モバイルデータトラフィックは時間帯や地域によって偏りが発生することが一般的に知られており[2][3], モバイルデータ通信における帯域利用効率が低下する課題がある．帯域利用効率を最大化する方法として，携帯電話基地局（eNB：evolved Node B）負荷を分散するために携帯端末（UE：User Equipment）の送信レートを制御する方法が考えられる．例えば，ある程度の遅延を許容し，通信の即時性を求めないデータ[4]（以下，遅延耐性データ）に着目して UE の送信レートを制御し，帯域利用効率を向上させることを目的とした Mobile Data Offloading Protocol（MDOP）が提案されている[5]．

1 静岡大学大学院総合科学技術研究科
Graduate School of Integrated Science and Technology, Shizuoka University
2 静岡大学情報学部
Facility of Informatics, Shizuoka University
3 静岡大学学術院情報学領域
College of Informatics, Academic Institute, Shizuoka University

MDOP は時間的，空間的，通信路的の三つの手法で eNB 負荷を分散させるプロトコルである．しかし， MDOP の時間的オフローディングにおいて，eNB 負荷や UE が生成するトラフィック，トポロジ構造などの多種多様なモバイルデータトラフィックの特性に対して，定式化した送信レート制御モデルで常に帯域利用効率を最大化するように送信レート制御を実施することは困難であり，送信レート制御手法にはまだ検討の余地がある．一方で，近年強化学習が様々な分野で応用されており，コンピュータゲームや囲碁ではプロに打ち勝つといった事例がある[6,7]. その他にも，ロボティクスや税金の管理システムに適用されるなど，強化学習アルゴリズムが注目を集めている[8-10]．今後更に様々な分野で成功事例が増えてゆくことが見込まれる．

本研究では，MDOP の時間的オフローディングにおける帯域利用効率を最大化するため，分散深層強化学習を用いたモバイルデータオフローディング手法を提案する．モバイルデータトラフィックの特性に基づき，帯域利用効率を最大化するような送信レートを学習することで，適切な送信量と送信タイミングを制御可能になると考える．また，複数 UE のデータを共有する分散学習で送信レート制御モデルを構築することで，学習モデルが単一の UE では経験できない状態を学習可能となり，汎用性の高い送信レート制御モデルを構築できる．

以下，本稿の構成を示す．第 2 章で関連研究について述べ，第 3 章で分散深層強化学習を用いたモバイルデータオフローディング手法について提案する．第 4 章で提案手法の評価結果を述べ，第 5 章で本稿をまとめる．

2. 関連研究

送信レート制御によって，帯域利用効率を高める既存研究として様々な手法が提案されている［11-13］．User Plane Congestion Management（UPCON）[11]は，通信設備の負荷状況やコンテンツの種類，ユーザの契約状況等を考慮して Quality of Service（QoS）制御することで通信インフラへの負荷を分散させる手法である．実際に UPCON を模擬した手法[12]では，UPCON が目標とするデータ到達率の改善に有効であることが示されている．また，ビデオデータなどの短い遅延を許容するデータを要求する各 UE に対して等しい帯域を割り当てる手法[13]が提案されているが，[12]は QoS の状態，[13]は多様なコンテンツが混在する環境下で帯域利用効率に偏りが生じてしまう．また，コンテンツの遅延耐性を考慮していないため，更なる帯域利用効率の向上が期待できると考える．

一方，ネットワークを最適化し，帯域利用効率を高める研究も提案されている．ネットワーク品質の向上を目的とした手法として，モバイルデータトラフィックの動的な変化に応じて，最適なネットワークを自動的に構築する Self-Organizing-Network（SON）[14]がある．SON は LTE ネットワークを対象にネットワークを自動設定，自動最適化，自動修復，自律計画する特徴があり，四つの機能でネットワーク品質の向上を図る．モバイルデータ通信に強化学習を適用した既存研究として，強化学習の一つである Q 学習[15]を，マクロセルとフェムトセルが混在するヘテロジニアスネットワークのチャネル選択に適用した手法[16]がある．評価の結果から，チャネル選択を人手で設定した規則に従って行う手法を，強化学習を適用したチャネル選択手法に変更することで，フェムトセル間の干渉が緩和されることを示している．しかし，[14][16]は送信レート制御による eNB 負荷の平滑化を行わないため，モバイルデータトラフィックの局所性を解消することは困難である．

既存研究の送信レート制御モデルやチャネル選択手法は，人手で構築された数理モデルでネットワークを最適化する手法であり，多種多様な通信インフラの状況に応じて，常に帯域利用効率を最大化することは困難である．また，今後 IoT が普及し，現在のモバイルデータトラフィックとは異なる傾向に変化した場合，一度定式化した数理モデルではモバイルデータオフローディングの性能に限界がある．

本研究では，人手で設定した規則に従って帯域を割り当てる送信レート制御に対し分散深層強化学習を適用する．送信レート制御モデルが学習を繰り返すことで，定式化された送信レート制御モデルでは追従できない突発的なモバイルデータトラフィックの変動や，トラフィックの推移を追従可能となり，多種多様な通信インフラの負荷状況に対して，帯域利用効率を向上できると考える．また，複数 UE のデータを共有する分散学習でモデル構築することで，学習モデルが単一の UE では経験できない状態を学習可能となり，汎用性の高い送信レート制御モデルを構築できる．

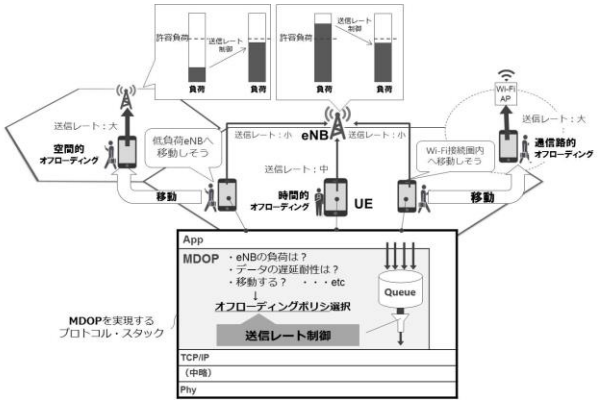

図 1 MDOP の概要[5]

3. 提案手法

3.1 概要

多種多様なモバイルデータトラフィックの特性に対して帯域利用効率を最大化するため，分散深層強化学習を用いたモバイルデータオフローディング手法を提案する．これまでに研究開発を進めてきた MDOP の時間的オフローディングに焦点を当て強化学習の適用可能性について検討を行う．**図 1** に MDOP の概要を示す．MDOP はある程度の遅延を許容する即時性を求めないモバイルデータトラフィックに対して，UE の送信レートを制御することで eNB への負荷を分散させ，帯域利用効率を向上させる． MDOP の送信レート制御方法には，Wi-Fi や LTE，5G などの異なるモバイル通信路に切り替えることで，トラフィックを分散させ，モバイルデータ通信路上のトラフィックを削減する「通信路的オフローディング」，UE の移動経路から低負荷な eNB を導出し，低負荷 eNB で通信するように促すことで，空間的局所性を解消する「空間的オフローディング」，eNB 負荷の時間的変動に応じて送信レートを制御し，eNB 負荷が低負荷である時間帯に通信を行うことで時間的局所性を解消する「時間的オフローディング」の 3 つがある． 3 つの制御方法の中から UE と eNB の状態やコンテンツの遅延耐性に応じて通信路・空間・時間の順にオフローディング条件を適用し，送信レート制御方法を決定する．ただし，どのオフローディング方法を適用した場合でも，コンテンツの遅延耐性時間がオフローディング方法で定められたデータ到達時間を超過する場合や，いずれのオフローディング方法にも該当しない場合は，オフローディングを行わず直ちに最大送信レートでデータを送信する．

MDOP の時間的オフローディングではモバイルデータトラフィックの特性に対し常に帯域利用効率を最大化するような送信レート制御手法はまだ検討されていない．そこで，

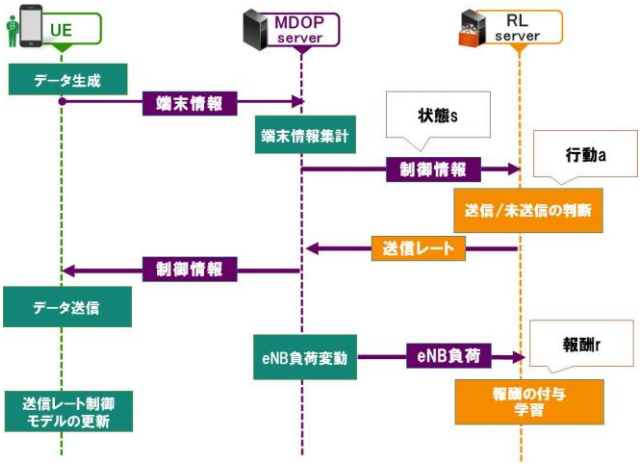

図 2 分散深層強化学習を用いた通信フロー

分散深層強化学習を用いて帯域利用効率の最大化を図る．**図 2** に分散深層強化学習を適用した時間的オフローディングにおけるアップロード時の通信フローを示す． UE が遅延耐性を持つアプリケーションを一時的に蓄積し，通信インフラの状況に応じて UE と MDOP サーバが送受信する制御情報を元に送信レートを制御することで，モバイルデータオフローディングを行う．端末情報には UE が保持しているコンテンツ量や接続先 eNB 情報，制御情報には eNB の負荷情報や eNB 接続ユーザ数などの送信レートを決定するために必要な情報が含まれる．

3.2 送信レート制御手法

定式化された送信レート制御モデルでは追従できないモバイルデータトラフィックの変動を追従可能にするため，送信レート制御手法に分散深層強化学習を適用する．また，学習を繰り返すことで，多種多様な通信インフラの負荷状況に対して，常に帯域利用効率を最大限利用可能とする送信レート制御モデルの構築を行う．提案手法では，試行錯誤しながら行動を最適化する強化学習の一つである Q 学習を用いる．Q 学習は，ある状態sにおいて取りうる行動aの価値を行動価値関数$Q(s,a)$として定量化し，試行錯誤しながら$Q(s,a)$を最大化するように逐次更新することで，各状態における行動を最適化する．式（1）に Q 学習における$Q(s,a)$の更新式を示す．

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right) \qquad (1)$$

$Q(s,a)$の更新は，全ての状態sと行動aの組に対して$Q(s,a)$を作成し，全ての$Q(s,a)$を任意の値に初期化した後，式(1)を用いて更新し学習する．したがって，ある問題に対して Q 学習を適用するには，起こりうる状態sと，状態sで取りうる行動aの組み合わせ$Q(s,a)$を事前に用意する必要がある．しかし，多種多様なモバイルデータトラフィックの状況に対して$Q(s,a)$を作成する場合，状態数が高次元なものとなるため事前に$Q(s,a)$を用意するのは困難である．暫定的に定めて学習を行ったとしても膨大な学習時間が必要となり，学習が収束しないおそれがある．そこで，$Q(s,a)$をニューラルネットワークで近似した Deep Q-network（DQN）[6]を適用する．$Q(s,a)$をニューラルネットワークで近似する手法は以前から提案されていたが，ニューラルネットワークで$Q(s,a)$を近似した場合，パラメータが増加し，学習が発散することが知られていた．学習が発散する要因としてデータ間の相関が高いこと，$Q(s,a)$の更新が行動選択の戦略を大きく変えてしまうことなどが挙げられる．DQN では，過去の状態や行動を保持しておき，学習時には保持した情報をランダムサンプリングして利用することで使用するサンプルの偏りを抑制する Experience Replay や，近似するべきパラメータを固定して回帰問題を解く fittedQ にニューラルネットワークを利用した neural fitted Q を利用することで学習則を安定化し，学習が発散することを回避している．そのため，DQN を適用することで，モバイルデータトラフィックのような高次元なデータに対しても，学習を収束させることができる．しかし，DQN は，行動 a を選択するモデルと評価するモデルが同じモデルであるため，行動 a を過大評価してしまい，精度が落ちてしまうおそれがある．そこで，行動選択時と評価時で異なるモデルを使用する Double Deep Q-network（DDQN）[17]を適用する．

提案手法では，図 2 中の Reinforcement Learning Server（RL サーバ）が DDQN を用いて送信レートを決定する． RL サーバは MDOP サーバから送られる制御情報を状態とし，状態から帯域利用効率を最大化するように送信レートを決定し，決定した送信レートで UE がデータを送信する．適切な送信レートが未知の場合，RL サーバが帯域利用効率を最大化するのは困難であるため，RL サーバが送信レート決定時に適切であると判断した送信レートを暫定的に最適な送信レートとする． RL サーバの報酬は UE がデータ送信後の eNB 負荷状況の変化から，送信レートが帯域利用効率を最大化する適切な送信レートであるか評価し付与する．報酬を付与した後に再び送信レートを決定する場合，RL サーバが得られた報酬を元に更新した送信レート制御モデルで送信レートを決定するので，状態から送信レートを決定し報酬を付与する過程を繰り返し学習することで，RL サーバが帯域利用効率を最大化する送信レートを獲得する．

3.3 分散深層強化学習

強化学習の学習過程において，常に現時点で最善であると思われる行動を選択すると，局所解に陥る可能性がある．そのため，行動の選択に ε の確率でランダムな行動を選択し，(1-ε)の確率で最大の Q 値を選択する ε-greedy 法が一般的に用いられるが，ε-greedy による行動の選択や DDQN のニューラルネットワークの各重みの初期値のランダム性によって同じ学習データ，同じネットワーク構造で学習を実

表 1 通信環境モデル

項目	設定値
UE 送信電力	23dBm
eNB 送信電力	46dBm
ISD（Inter side distance）	500m
帯域幅	10MHz
周波数	2.0GHz
TCP	New Reno

表 2 学習用パラメータ

項目			設定値
状態 s			遅延不可データ
			1 秒前の遅延不可データ
			理想負荷
			接続 UE 数
行動 a			最大送信量送信（0~160）
報酬 r	$L_{abs} \leq 500$		+1
	$L_{abs} > 500$	送信	-1
		未送信	+1
学習率 α			0.001
割引率 γ			0.9
ε -greedy			0.01

施した場合でも，モデルの精度にばらつきが生じるおそれがある．そこで，ランダム性によるモデル精度のばらつきを解消するために, 同じ DDQN の構造を持つモデルを予め複数用意し，同じ学習データを分散学習させ，最終的にモデルの統合する分散深層強化学習を行う．複数のモデルを統合させることで，ε-greedy やニューラルネットワークの重みの初期値によるランダム性に依存しない分散送信レート制御モデルとなる．さらに，異なる学習データで構築したモデルを統合することで，汎用性の高い送信レート制御モデルとなる．MDOP の場合，複数 UE でモデル構築を行うことで，単一の UE のみでは得られない UE トラフィックの特性の学習が期待できる．

4. 評価

4.1 評価方針

提案手法が帯域利用効率を向上させることを確認するため，eNB の負荷変動に応じた送信レート制御を行い，時間的局所性を解消できるかシミュレーションで評価を行った．MDOP の時間的オフローディングは UE が移動しない環境を想定しているため，UE は移動しないものとする．通信環境モデルは 3GPP が推奨するモデル［18］に基づき作成した．**表 1** に通信環境モデルを示す．シナリオを作成，評価するにあたり，LTE 環境を詳細に再現できるネットワークシミュレータ Scenargie［19］を用いた．

表 3 ネットワーク構造

層種	ユニット数	活性化関数
fc1	6	Tanh
fc2	100	Tanh
fc3	200	Tanh
fc4	400	Tanh
fc5	161	-

4.2 ネットワーク構成

分散深層強化学習を用いた送信レート制御の評価を行うため，DDQN を適用する．はじめに，DDQN を適用するにあたり，学習時に必要となる各種パラメータを定義する．**表 2** に Q 学習の各種パラメータを示す．状態 s は図 2 の通信フローで MDOP サーバから受け取る制御情報に加え, 時刻 t[s]において，時刻 t-1[s]の遅延できないデータ（以下遅延不可データ）を状態 s とする．MDOP サーバが送る制御情報は送信レートを決定するために必要な情報であるため, 制御情報のみで送信レートの算出は可能であるが，1 秒前の eNB の負荷情報を加えることで, DDQN のネットワークが，モバイルデータトラフィックの時系列変化を考慮した学習を行うと期待できる．行動 a は実際に UE がデータ送信する送信量とし，最大送信量を 16000[byte/sec]とする．しかし，行動 a を送信量とした場合，状態数が 16000 となるため状態数が膨大となる．状態数が高次元の場合，学習が収束しないおそれがあるため，行動 a の状態数を 160 とし，実際に UE がデータ送信する送信量は，行動 a の状態を 100 倍した送信量にすることで，行動 a の状態数を削減する．行動 a の状態数を削減するとネットワークの規模が小さくなるので，学習の収束が期待できる．報酬 r は送信レート制御することで目標とする eNB 負荷（以下，理想負荷）にどれほど近づけることができたかに対して，その重みを動的に変更するべきであるが，学習を高速化するために報酬の Cliping を行う．報酬 r は+1,-1 の報酬を与えることとし，報酬を付与するか否かの判断に，送信レート制御後の eNB 負荷と理想負荷の絶対値誤差L_{abs}を用いる．L_{abs}が一定の範囲内であれば+1，範囲外であれば-1 の報酬を与えることとし，L_{abs}が 500byte/sec の範囲内外であるかを閾値として報酬を付与する．ただし，現状負荷が理想負荷を超過する場合に L_{abs}が 500byte/sec を上回った場合，データ送信を行わなければ適切な送信レート制御をしたと言えるため，+1 の報酬を与える．

DDQN のネットワークは，入力は状態 s，出力は Q(s,a)

表 4 通信環境モデル

項目	設定値
シミュレーション時間	1800s
UE 数	1 台
UE の最大送信レート	16000[byte/sec]
データの種類	FTP
データ生成時間	0s~1800s
eNB 数	1 台
eNB の最大受信量	20000[byte/sec]
理想負荷	16000[byte/sec]

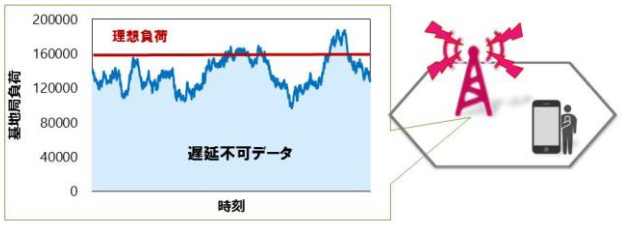

図 3 評価トポロジ

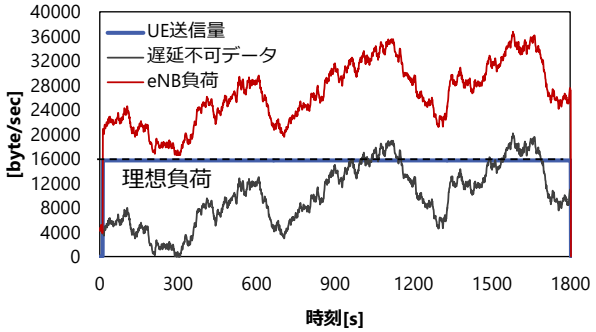

図 4 送信レート制御なしの eNB 負荷と UE の送信量

のニューラルネットワークで構成される. **表 3** に DDQN のネットワーク構造を示す．DDQN のネットワーク構造は，5 層の全結合層で構成されており，活性化関数は tanh を使用する．また，学習の過程において，常に現時点で最善であると思われる行動を選択すると，局所解に陥る可能性がある．そのため，行動の選択に ε の確率でランダムな行動を選択し，(1-ε)の確率で最大の Q 値を選択する ε-greedy 法を採用することで，局所解に陥ることを回避する．特に，ε-greedy 法の中でも．学習回数の増加と共に ε の値を徐々に減らしていく LinearDecayEpsilonGreedy を使用する．

4.3 MDOP への適用評価

MDOP への適用評価では，eNB が時間経過に伴い変化するシナリオを想定し，シナリオを提案手法で学習させた結果，学習を繰り返すことで，eNB 負荷変動に応じた送信レート制御ができるか評価した．また，4.2 節ネットワーク構成で，状態 s に 1 秒前の遅延不可データを採用すること

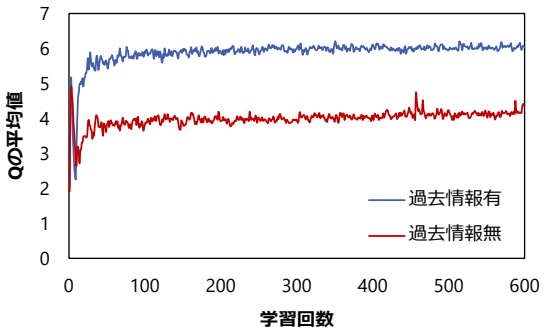

図 5 過去情報有・無の Q の平均値

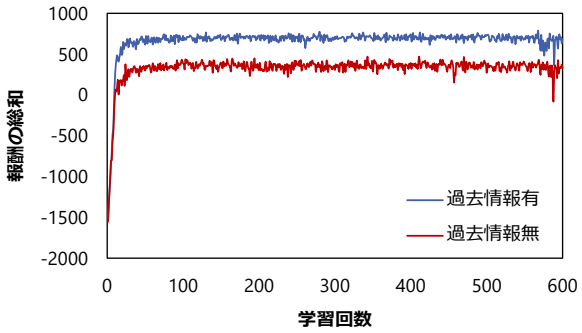

図 6 過去情報有・無の報酬の総和

で，モバイルデータトラフィックの時系列変化を考慮したネットワーク構造とした．過去情報が帯域利用効率の精度にどのような影響を及ぼすか検証するため，1 秒前の遅延不可データを状態 s に考慮したモデル（以下，過去情報有モデル）と考慮しないモデル（以下，過去情報無モデル）を同一シナリオに適用し帯域利用効率の比較を行った．**図 3** に評価トポロジ，**表 4** に評価シナリオをそれぞれ示す．

図 4 に送信レート制御なしの場合の eNB 負荷と UE の送信量を示す．制御なしの場合，常に UE が最大送信レートでデータ送信をしており，eNB 負荷が理想負荷を超過していることが確認できる．**図 5** に過去情報有モデルと過去情報無モデル（以下，過去情報有・無モデル）のシミュレーション毎の Q 値の平均値，**図 6** に過去情報有・無モデルの報酬の総和をそれぞれ示す．過去情報有・無モデルの Q 値の平均値と報酬の総和がシミュレーション回数に比例して増加・収束していることから，提案手法がシミュレーションを繰り返すことで eNB 負荷に適切な送信レートを学習し，学習が収束していることがわかる．過去情報有・無モデルの帯域利用効率の比較には，報酬の総和が最も高いシミュレーション回数の送信レートモデルを採用する．過去情報有モデルは 568 回目，過去情報無モデルは 343 回目の送信レート制御モデルをそれぞれ使用して送信レート制御を行い，提案手法が eNB 負荷変動に応じた送信レート制御ができるか検証する．**図 7** に過去情報有・無モデルの eNB 負荷，**図 8** に過去情報有・無モデルの UE の送信量を示す．遅延不可データと UE の送信量を比較すると，図 8 から UE

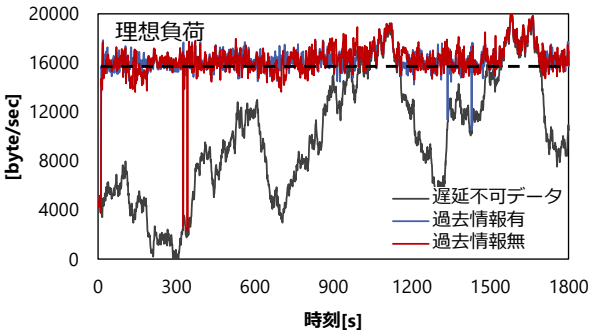

図 7 過去情報有・無の eNB 負荷

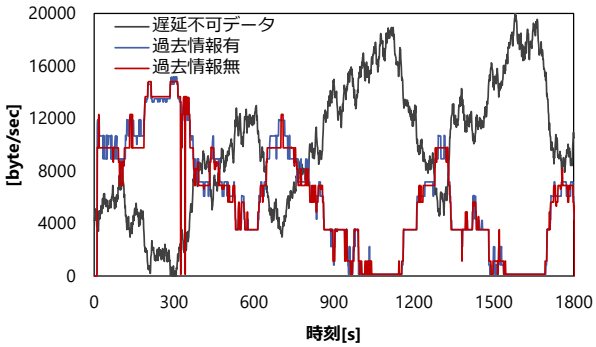

図 8 過去情報有・無の UE 送信量

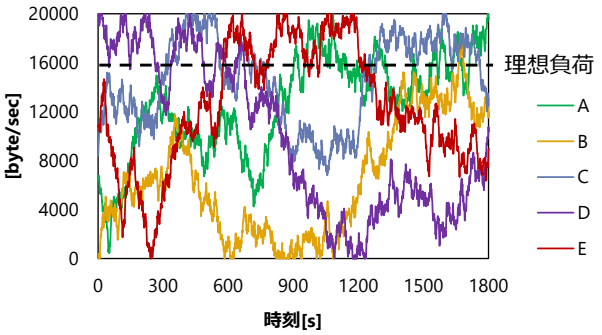

図 9 学習データ

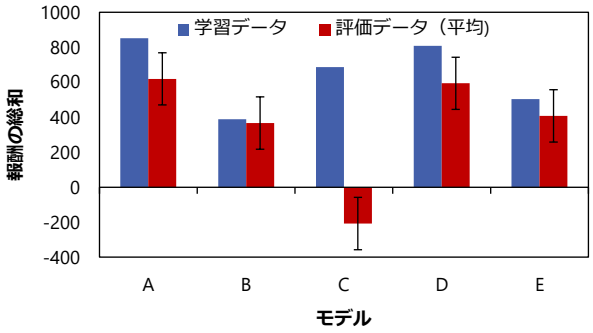

図 10 入力データ別の報酬の総和

が eNB 負荷を理想負荷に近づけるように送信レートを制御していることに加え，理想負荷を超過する区間では送信していない．そのため，提案手法が eNB 負荷変動に応じた送信レート制御を学習し，理想負荷に近づけるように制御したと考えられる．また，図 7 から過去情報有・無モデルの eNB 負荷を比較すると，過去情報有モデルが過去情報無モデルに比べて eNB 負荷を理想負荷に近似している．過去情報を考慮しない場合，送信レート制御モデルは現在の eNB 負荷から適切な送信レートを算出するが，状態 s に 1 秒前 の遅延不可データを採用することで，ニューラルネットワークが過去，現在の状態から次にどのような状態に遷移するかを推測することが可能となる．過去情報有モデルのネットワーク構造では送信レート制御時と 1 秒前の eNB 負荷情報を保持するため，二点の負荷変動から次にどのようなトラフィック変動になりうるかを線形回帰で予測可能となったと考える．トラフィックの時間的推移を考慮した上で適切な送信レート制御をすることを可能とし，帯域利用効率を向上させたと考える．一方，1200s～1500s の区間では過去情報有モデルの eNB 負荷が振動している．原因として，送信レート制御するために必要な次元数が現在の入力次元数では不足しており，膨大なトラフィックパターンを網羅する事が困難であったと考える．そのため，学習パラメータやネットワーク構造を再検討することで更なる精度向上が見込まれる．以上の結果から，提案手法が時間的局所性を解消し，帯域利用効率が向上することを確認した．

4.4 トラフィック別の帯域利用効率の評価

トラフィック別の帯域利用効率の評価では，送信レート制御モデルを構築時に使用する遅延不可データ（以下，学習データ）とは異なる時間的局所性が発生するシナリオを複数用意し，提案手法が学習データと異なるモバイルデータトラフィックに対して時間的局所性を解消する送信レート制御を行うことができるか評価を行った．また，学習データによって，帯域利用効率の精度がどのように変化するか評価した．通信環境モデルや評価トポロジは MDOP への適用評価と同一のものを使用した．学習データと構築した送信レート制御モデルの帯域利用効率を評価する際に使用するデータ（以下，評価データ）を用意し，データ数は学習データ 5 件，評価データ 50 件とした．帯域利用効率を評価する指標として，送信レートが帯域利用効率を最大化する適切な送信レートであるか判断し付与する報酬の総和を採用し，各評価データで得られた報酬の総和の平均値を比較することで，帯域利用効率の精度を評価した．

図 9 に学習データ，**図 10** に送信レート制御モデル別の学習データ適用時の報酬の総和と評価データ適用時の報酬の総和の平均値をそれぞれ示す．図 10 から，学習データが異なる場合，同じ評価データに対しても報酬の総和の平均値が異なり，学習データの報酬の総和に比べ，評価データの報酬の総和の平均値は低いことが確認できる．モデル C 以外の送信レート制御モデルは，評価データに適用した場合でも正の報酬を獲得している．モデル C は，図 9 の学習データが，大半の期間で理想負荷を上回るモバイルデータト

ラフィックであるため，eNB 負荷が理想負荷を超過する場合には送信するべきでない事を学習はできるが，理想負荷を下回る場合，どのような送信レート制御をすれば適切であるのかを学習できず，評価データに対する報酬の総和の平均値が低くなったと考える．正の報酬を獲得しているモデルの中でも，報酬の総和が最も高いモデル A と報酬の総和が最も低いモデル B の報酬の総和の平均値はそれぞれ 619,366 であり，正の報酬が付与されるL_{abs}の範囲内にモバイルデータトラフィック全体の 67%，62%が収まる結果となった．図 9 の学習データは，モデル A,D のように時間推移に伴い緩やかに増加・減少傾向となるトラフィックと，モデル B,E のように上下に激しく振動するトラフィックがあり，入力次元に過去情報を採用することで二点の負荷変動から線形回帰で予測可能となったとしても，A や D のように，緩やかに変化するデータに対しては追従が容易であるが，B,E 大きく振動するデータに対しては追従が困難である．そのため，学習データの傾向の違いが，モデルの帯域利用効率に大きく影響したと言える．本評価に使用した送信レート制御モデルは，図 9 の学習データのみを学習したモデルであるため，学習データ以外のモバイルデータトラフィックの特性を繰り返し学習させることで，未知のモバイルデータトラフィックに対して更に帯域利用効率を向上させることを期待できる．以上の結果から，学習時に使用したシナリオと帯域利用効率を比較すると帯域利用効率が低下するが，未知のモバイルデータトラフィックに対しても，送信レート制御モデルが時間的局所性を解消することを確認した．

4.5 分散送信レート制御モデルの評価

分散送信レート制御モデルの評価では，複数 UE がそれぞれ送信レート制御モデルを構築し，送信レート制御モデルを統合した分散送信レート制御モデル（以下，分散モデル）と，単一の送信レート制御モデル（以下，単一モデル）を同一のシナリオに適用，評価することで，単一モデルで誤って判断した送信レート制御に対しても，分散モデルが適切に送信レート制御し，ニューラルネットワークの重みの初期値や ε-greedy 法のランダム性による帯域利用効率の影響を軽減できるか評価を行った．MDOP への適用評価と同じネットワーク構造，学習データを用いて学習，評価を行った．分散モデルの送信レートの決定は，各モデルが決定した送信レートの平均値を採用する．

単一モデル別の帯域利用効率と分散モデルの統合モデル数別の帯域利用効率の比較を行った．統合モデル数を 10 とし，報酬の総和が高いモデルから順に統合，評価を行う．帯域利用効率の評価指標として，報酬の総和を用いた．**図 11** に単一モデル別の報酬の総和を示す．報酬の総和が最も高いモデル A と最も低いモデル J の報酬の総和がそれぞれ 724,438 であり，同じ学習データ，ネットワーク構造で学習した場合でも，ニューラルネットワークの重みの初期値

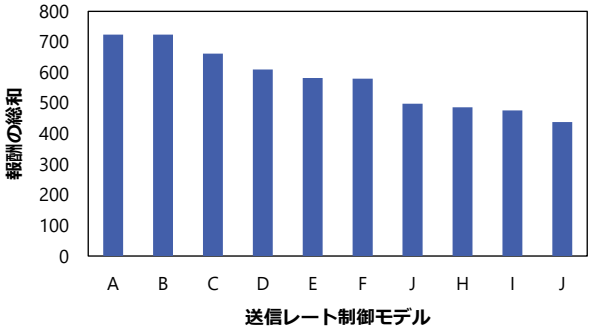

図 11 送信レート制御モデル別の報酬の総和

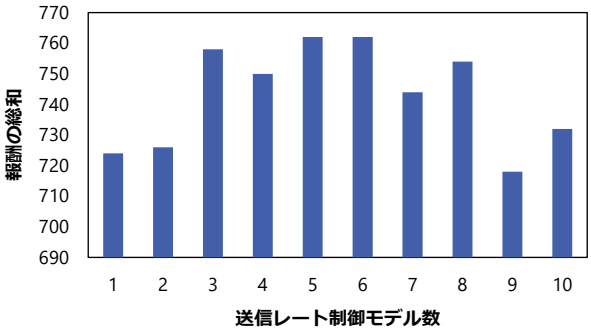

図 12 統合モデル数別の報酬の総和

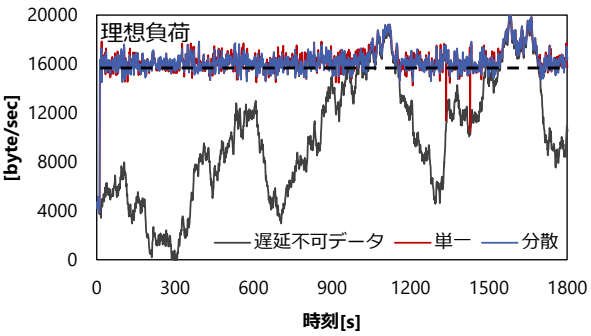

図 13 単一・分散モデルの eNB 負荷

や ε-greedy 法のランダム性が，モデルの帯域利用効率に影響を及ぼすことが確認できる．統合モデル数別の帯域利用効率の比較には，報酬の総和が高い A から順にモデルを統合した．**図 12** に統合モデル数別の報酬の総和を示す．図 12 から，統合モデル数 5 までは，統合モデル数の増加に比例して報酬の総和が高くなるが，統合モデル数 5 以降は減少傾向にある．統合モデル数が増加することで，単一モデルが誤って判断した送信レートを補正することが可能となり，帯域利用効率が向上するが，図 11 から各モデルは帯域利用効率の精度に差があるため，統合モデル数が増加し，分散モデルの規模が一定数を超過すると，不適切な送信レート制御を行うモデルが多くなるため，帯域利用効率が低下したと考える．

分散モデルと単一モデルの帯域利用効率の比較には，報酬の総和が最も高い統合モデル数 5 の分散モデルを用いて

評価した．**図 13** に単一モデルと分散モデルを適用した時の eNB 負荷を示す．図 13 の結果から，1200s~1500s の区間で，単一モデルでは eNB 負荷が振動しているが，分散送信レート制御モデルでは eNB 負荷の振動が解消されていることが確認できる．報酬の総和は単一モデルが 724，分散モデルが 767 であり，分散モデルを適用することで帯域利用効率が 6%向上した．分散モデルは複数のモデルが決定した送信レートの平均値を送信レートとするため，単一モデルが持つ不安定な送信レート制御部分を補正できたためである．一方で，eNB 負荷が振動しない区間では，単一モデルと分散モデルの帯域利用効率に大きな変化は見られなかった．本評価では複数のモデルが同一の学習データで学習しているため，送信レートの補正はできるが，更に微細な送信レート制御は困難であったと考える．したがって，報酬付与の指標であるL_{abs}の範囲が更に狭い送信レート制御モデルを組み合わせることで，帯域利用効率の向上が見込める．また，異なるモバイルデータトラフィックの特性を学習したモデルを統合することで，未知のモバイルデータトラフィックに対して更なる帯域利用効率の向上が期待できる．以上の結果から，分散モデルが，単一モデルに比べ帯域利用効率を向上させることを確認した．

5. おわりに

MDOP の時間的オフローディングにおける帯域利用効率の向上を目的とした，分散深層強化学習を用いたモバイルデータオフローディング手法を提案した．MDOP への適用評価，トラフィック別の帯域利用効率の評価の結果，提案手法を用いることでシミュレーション回数の増加に伴い帯域利用効率が向上し，時間的局所性を解消することを確認した．また，未知のモバイルデータトラフィックに対しても，時間的局所性を解消するような送信レート制御を行うことを確認した．分散送信レート制御モデルの評価では，単一モデルでは不安定な送信レート制御を行う区間で適切な送信レート制御をすることができた．したがって，提案手法が帯域利用効率を向上したといえる．

今後，UE や eNB を増加させ実環境に基づいた評価を行う予定である．また，ネットワーク構造を再検討し，更なる帯域利用効率の向上を目指す．

謝辞

本研究は，科学研究費補助金基盤研究（B）「深層強化学習を用いたモバイルデータ 3D オフローディングの研究（17H01730）」の支援を受けたものである．

参考文献

[1] Cisco Visual Networking Index.: Global Mobile Data Traffic Forecast Update, 2016-2021. White Paper (2017).
[2] 総務省：我が国の移動通信トラヒックの現状 (2015).
[3] NTT ドコモ：電波政策ビジョン懇親会ヒアリング資料 (2014).
[4] A.Biral et al.:The challengesof M2M massive access in wireless cellular networks, Digital Communications and Networks, 1.1, pp.1-19(2015).
[5] 西岡哲朗，ほか：モバイルデータトラフィックの時間的局所性を解消するモバイルデータオフローディングプロトコルの提案，情報処理学会論文誌 Vol.58, No.1, pp.13-23(2017).
[6] V.Mnih, et al.: Human-level control through deep reinforcement learning. Nature 518.7540 ,pp.529-533(2015).
[7] D.Silver et al.; Mastering the game of Go with deep neural networks and tree search. Nature 529.7587 pp.484-489(2016).
[8] L.Busoniu, et al.: A comprehensive survey of multiagent reinforcement learning. IEEE Transactions on Systems Man and Cybernetics Part C Applications and Reviews 38.2 :156(2008).
[9] 保田俊行，ほか： 実例に基づく強化学習法の頑健性向上に関する一考察. 計測自動制御学会論文集 pp1150-1157(2016).
[10] N.Abe, et al.: Optimizing debt collections using constrained reinforcement learning." Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM(2010).
[11] 3GPP TR23.705.:User Plane Congestion management (Release-12).
[12] 鈴木理基，ほか：LTE 網におけるサービス単位のトラヒック収容技術の検討，DICOMO2014 シンポジウム論文集，pp. 1326-1333(2014).
[13] Y.Timmer, et al.: Network Assisted Rate Adaptation for Conversational Video over LTE, Concept and Performance Evaluation, Proceedings of the 2014 ACM SIGCOMM workshop on Capacity sharing workshop, pp.45-50(2014).
[14] 3GPP TR 36.902.:Evolved Universal Terrestrial Radio Access Network(E-UTRAN); use cases and solutions (Rel-9)(2009).
[15] C.J.C.H. Watkins :Learning from Delayed Rewards, Cambridge University PhD thesis(1989).
[16] M.Bennis, et al.: A Q-learning Based Approach to Interference Avoidance in Self-Organized Femtocell Networks, Globecom Workshops, pp.706-710 (2010).
[17] H.Van Hasselt, et al.: Deep Reinforcement Learning with Double Q-Learning. AAAI. (2016).
[18] 3GPP TR 25.942: Radio Frequency (RF) system scenarios(Rel-13) (2015).
[19] Space-Time Engineering, LLC.: Scenargie, https://www.spacetime-eng.com/en/products,(2017).