Vol.2018-ARC-231 No.10
2018/6/15

強化学習を用いた無線センサネットワークのデータ通信経路自律制御の検討

呉 天雄[1,a)]　近藤 正章[1]

概要：マルチホップ無線センサネットワークは近年広く用いられるようになってきている．センサノードの多くはバッテリー駆動であり，センサノードの消費電力を抑えつつ，バッテリー切れでダウンするセンサノードの発生をなるべく遅くするようにセンサネットワークを制御することは，無線センサネットワークの重要な課題である．特に，各センサノード間のエネルギー残量やセンシング間隔が異なるときに，センサネットワークのデータ通信経路を制御することは難しい問題となる．本稿では，強化学習手法を用いて自律的に最良と考えられる無線センサネットワークデータ通信経路を求める手法を提案する．提案手法の検討にあたり，センサネットワーク全体を管理できるマスタ計算機の存在を仮定した問題設定から検討を始め，より現実的な状況として，各センサノードがネットワークの一部のノードの情報のみを持っている状態で通信経路の自律最適化を行う手法も検討する．初期評価として格子状のセンサネットワークモデルを導入し，シミュレーションにより提案手法の評価を行った．評価結果より，固定化されたデータ通信経路を用いる場合よりも，提案した強化学習による自律制御を用いる場合のほうがセンサネットワーク全体のライフタイムを長くすることができることが確認できた．

1. はじめに

センサネットワークはセンサをネットワークで相互に接続することにより，多数のセンサによって得られた情報を収集・利用するシステムである．無線センサネットワーク (Wireless Sensor Network: WSN) は，通常はバッテリ駆動のセンサから無線通信でデータを収集することで，インフラ整備の必要がなく低コストで導入できるため，部屋やビルの中で用いられるような生活空間規模のものから，街の中や農地・砂漠など地域規模，地球規模といった様々なシーンで用いられるようになってきた．

WSN では，センサノードから得られたセンシング情報を基地局となるシンクノードまで伝達する必要がある．WSN においてセンシング情報をシンクノードに送信する方法は大きく 2 つあり，各センサノードが情報を直接シンクノードに通信するシングルホップ WSN と，センサノードを中継しながら通信を行うマルチホップ WSN があげられる．その中で，マルチホップ WSN は，シングルホップ WSN に比べ，通信効率やエネルギー効率に優れており，ネットワークの拡張性が高いため，特に注目されている手法である．

WSN のセンサノードはバッテリ駆動の場合が多いため，いかに個々のセンサの電力消費を低く抑えるかや，センサネットワーク全体から情報を効率良く収集する上でエネルギー切れでダウンするセンサノードをいかに減らすかが重要となる．センサ間の通信に対する省電力技術としては，ノードがデータの送受信をしていないアイドル状態時の電力削減手法 [1] が提案されてきたが，その他にもデータ通信経路を最適化することも重要である．マルチホップ WSN の通信経路最適化に関する先行研究には，2 つの特定のノード間のマルチホップ通信に対しての研究 [2] や，WSN をクラスタに分けることによって省電力化を目指した研究 [3], [4] などがあるが，多数のセンサノードからシンクノードまでの通信経路の最適化を対象とする研究や，各センサノードのエネルギー残量やセンシング間隔を考慮しつつ自律的に制御を行う研究は少ない．

通信経路の最適化は，各センサノードのエネルギー残量やセンシング間隔が異なるときに，その解を解析的に求めることは難しく，また人手での経験的な通信経路の最適化も容易ではない．そこで，本稿ではセンサネットワークが人手を介さずに自律的に最良と考えられるデータ通信経路を求める手法として，強化学習を用いた自律制御について検討し，センサネットワーク全体のライフタイムを長くするような通信経路の制御手法を提案する．

まず，初期検討としてセンサネットワーク全体を管理で

[1] 東京大学大学院情報理工学系研究科
[a)] wu@hal.ipc.i.u-tokyo.ac.jp

きるマスタ計算機の存在を仮定した問題設定から検討を始める．一方で，センサノード全体の状況を常に把握し，個々のセンサノードに指示を出しながら経路制御を行うことは難しく，本問題設定では実用性が低いという問題がある．そこで，各センサノードがセンサネットワークの一部のみの情報に基き自律制御を行うマルチエージェント方式の強化学習手法を導入することで，現実的な手法構築を目指し検討を行う．

2. 強化学習の理論

2.1 強化学習の基礎

強化学習の枠組みはエージェントと環境の相互作用からなる．エージェントは行動決定の主体であり，環境はエージェントが相互作用を行う対象である．相互作用は状態・行動・報酬の 3 つの要素で構成される．状態は，エージェントが置かれている状況を表し，行動はエージェントが環境に対して行う働きかけを，報酬はエージェントが行った行動の即時的な良さを表す．エージェントが行動を決定するためのルールを方策とよび，強化学習の問題を解くということは，できるだけ多くの報酬が得られるように方策を設計することに相当する．しかし，すぐには小さな報酬しか得られない行動でも，後々に大きい報酬が得られる場合には，初めの行動は全体としてみれば良い行動であると言えるため，後から得られる報酬を考慮して行動を決定する必要がある．そこで，収益と呼ばれる，ある期間で得られた累積の報酬を表す指標を導入する．強化学習においてよく用いられる収益に割引報酬和があり，これは未来の不確実性を報酬を割り引く形で表した収益である．時間ステップ t で得られた報酬を R_t として，収益 G_t を

$$G_t = \sum_{\tau=0}^{\infty} \gamma^{\tau} R_{t+1+\tau} = R_{t+1} + \gamma R_{t+2} + \cdots \qquad (1)$$

と定義する．割引率 $\gamma (0 \leq \gamma \leq 1)$ はどれだけ未来を割り引くかを表した定数である．γ が 1 に近づくほど長期的に有益な行動を評価するようになり，γ が 0 に近づくほど，即時的に有益な行動を評価するようになる．

2.2 方策勾配に基づく強化学習アルゴリズム

エージェントの行動が確率的に表された確率的方策を考える．状態 s で行動 a を行う確率が，あるパラメータベクトル θ によって表された確率モデル $\pi_\theta(a|s)$ に従うとする．このアルゴリズムの目的は，パラメータ θ で表された期待収益 $J(\theta)$ を目的関数として，これを最大化するような確率的方策 π_θ のパラメータ θ を求める問題として捉えることができる．目的関数 $J(\theta)$ の最大化には勾配ベースの方法が用いられ，η を学習率として，

$$\theta \leftarrow \theta + \eta \frac{\partial}{\partial \theta} J(\theta) \qquad (2)$$

と繰り返し更新していくことで，局所最適解を求めることを目指す．

方策勾配に基づく強化学習アルゴリズムは以下の 3 ステップで表される．

1. 方策 π_θ による行動
2. 方策 π_θ の評価
3. 方策 π_θ の更新

以下の節では，それぞれのステップについて詳述する．

2.2.1 方策 π_θ による行動

状態空間 S が連続で，行動空間 A が離散である場合は，S が連続であることから，状態を数え上げることは不可能であるため，次のような線形関数を介したソフトマックス関数で確率的方策 π_θ を表すことを考える．

$$\pi_\theta(a|s) = \frac{\exp(\theta^{\mathrm{T}} \phi(s, a))}{\sum_{b \in A} \exp(\theta^{\mathrm{T}} \phi(s, b))} \qquad (3)$$

ここで，$\phi(s, a)$ は状態と行動によって定まるベクトルであり，θ が方策を調整するパラメータベクトルである．環境との相互作用を通じてパラメータ θ について更新することで学習を行う．

2.2.2 方策 π_θ の評価

方策勾配に基づく強化学習アルゴリズムでは，期待収益 $J(\theta)$ を目的関数として，これを最大化するようなパラメータ θ を求めることが目的であった．本稿では，式 (1) で表される割引報酬和を収益に用い，初期ステップからの割引報酬和の期待値を目的関数と考えると，

$$J(\theta;\, s_0) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} R_t \,\middle|\, S_0 = s_0 \right] \qquad (4)$$

とかける．ここで，S_0 はエージェントの初期状態を表す確率変数，s_0 はエージェントの初期状態の実現値を表す．

2.2.3 方策 π_θ の更新

式 (2) における $J(\theta)$ の勾配を求めることでパラメータ θ の更新を行う．割引報酬の期待値を目的関数としているため，方策 π に従って状態 s で行動 a を行ったときの行動価値を表す行動価値関数 $Q^\pi(s, a)$ は，

$$Q^\pi(s, a) = \mathbb{E}\left[\sum_{\tau=1}^{\infty} \gamma^{\tau-1} R_{t+\tau} \,\middle|\, s_t = s,\, a_t = a,\, \pi \right] \qquad (5)$$

で表され，$\frac{\partial}{\partial \theta} J(\theta)$ は，方策勾配定理 [6] により，行動価値関数 $Q(s, a)$ を用いて，

$$\frac{\partial}{\partial \theta} J(\theta) = \mathbb{E}_{\pi_\theta}\left[\frac{\partial}{\partial \theta} \pi_\theta(a|s) \frac{1}{\pi_\theta(a|s)} Q^\pi(s, a) \right] \qquad (6)$$

$$= \mathbb{E}_{\pi_\theta}\left[\frac{\partial}{\partial \theta} \log \pi_\theta(a|s) Q^\pi(s, a) \right] \qquad (7)$$

と表せる．

実際には，式 (7) で表される勾配を解析的に求めることはできないため，エージェントが確率的方策 π_θ に基づい

て行動を行い，得られたサンプルを利用して勾配の近似を行う．エージェントが方策 π_θ に従って T ステップの行動を M エピソードだけ行った結果を用いモンテカルロ近似により次のように近似する．

$$\frac{\partial}{\partial\theta}J(\theta) = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{T}\sum_{t=1}^{T}\frac{\partial}{\partial\theta}\log\pi_\theta(a_t^m|s_t^m)Q^\pi(s_t^m, a_t^m) \tag{8}$$

ここで，s_t^m と a_t^m はそれぞれ m エピソード目の t ステップ目の状態と行動である．

また，行動価値関数 $Q^\pi(s, a)$ も未知であるため，これを推定する必要がある．行動価値関数を即時報酬で近似し $Q^\pi(s, a) = R_t$ とすることで，以下の REINFORCE アルゴリズム [7] が導出される．

$$\frac{\partial}{\partial\theta}J(\theta) = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{T}\sum_{t=1}^{T}\frac{\partial}{\partial\theta}\log\pi_\theta(a_t^m|s_t^m)R_t^m \tag{9}$$

行動価値関数の近似法は，方策を表現する Actor と方策を評価する行動価値関数の 2 つをモデルとして保持する Actor-Critic アルゴリズム [6] や，確率分布間の擬距離を KL ダイバージェンスで定め自然勾配 [8] とよばれる勾配方向を用いてパラメータ θ の更新を行う自然方策勾配法 [9], [10] が他にあげられるが，本稿では初期検討として，REINFORCE アルゴリズムを用いた強化学習を利用する．

3. 提案手法

3.1 センサネットワークモデル

本稿では，強化学習を用いた通信経路の最適化を検討するにあたり，汎用的なセンサネットワークを考える前に，簡単化のためセンサノード 9 個とシンクノード 1 個を**図 1** のように格子点に配置した格子状ネットワークを想定する．格子は辺の長さを 1 単位とした正方形で構成されており，それぞれのセンサノードは左下より順に $S1$〜$S9$ と識別する．センサノードのエネルギー残量と保持しているデータ量を変数にとり，センサノード S_i のエネルギー残量と保持データ量をそれぞれ E_i, I_i とする．センサノードは定期的に情報収集を行うが，情報収集間隔はそれぞれ異なることも想定する．1 回のセンシングにより 1 単位の情報が得られ，その収集間隔を r_i とする．各センサノードは自分自身以外のノードにデータを送信でき，1 回の通信でその時点で持っている全てのデータを送信する．本稿では，上述のセンサネットワークモデルに対して，初期のエネルギー残量分布や保持データ量分布が与えられたときに，ネットワーク全体のライフタイムを長持ちさせるデータ通信経路を自律制御することが目的である．

3.2 問題設定

2 章で述べたセンサネットワークモデルに対して，方策勾配に基づく強化学習アルゴリズムを用いるための問題設定を行う．強化学習を行うエージェントは，センサネットワーク全体を管理できるデバイス 1 つの場合と，センサネットワークの各センサノードがそれぞれエージェントとなっている場合を考える．センサノード 9 個とシンクノード 1 個が図 1 のように格子点に配置されているとき，エージェントが取ることができる行動種別は，転送先ノード（シンクノードを含む）を選択して，当該ノードへデータを転送することである．通信距離 d，通信情報量 c のとき，送信側は $E_t cd^3$，受信側は $E_r cd^3$ のエネルギーを消費すると仮定する[*1]．ここで，E_t および E_r は，それぞれ単位情報量，単位距離あたりの通信エネルギーに相当する．

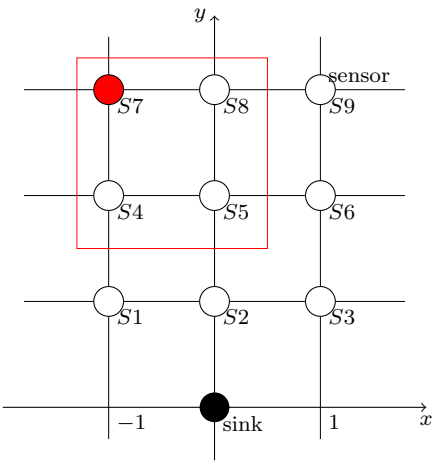

図 1 センサ配置とセンサノード番号，部分獲得情報の例

エージェントが観測できる環境は，マスタ計算機がセンサノード全体を管理しながら学習を行う場合は全てのセンサノードのエネルギー残量と保持データ量であるとし，各センサノードがマルチエージェントとして学習を行う場合は，自身とシンクノード側の最近接最大 3 ノードのエネルギー残量と保持データ量のみとする．例えば，センサ $S7$ が観測できるノードの範囲は図 1 に示したように，センサ $S4$, $S5$, $S7$, $S8$ となる．このとき，各センサノードはセンサネットワーク全体の状態に関する知識はなく，一部のセンサノードの状態のみが分かっているという条件のもとで，個々の学習結果に基づいて行動を選択する．

3.3 全体管理マスタ計算機方式

データ通信にかかる電力を削減しつつ，センサネットワーク全体のライフタイムを長くするためには，全ノードがなるべく均等に電力を消費しながら，かつ情報をできるだけシンクノードに近いノードに通信することが求められる．本節では，これを強化学習で達成するための報酬関数を全体を管理可能なマスタが存在するマスタ計算機方式について設計する．

本稿では，センサノードのエネルギー残量の差の最大値

[*1] 受信時も ACK の送信が必要なため，ここでは簡単のために受信側も距離の 3 乗に比例すると仮定した．

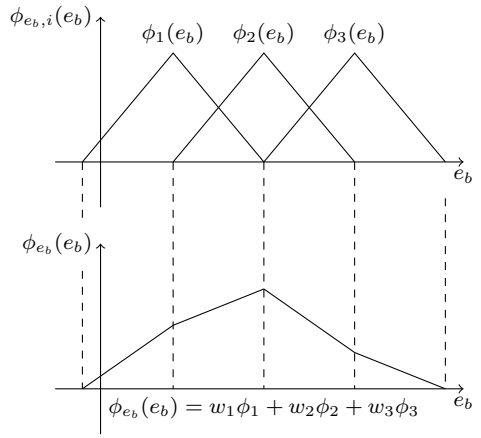

図 2 三角分布を用いた線形アーキテクチャ

を Δe_{old}，1 ステップ後の同じものを Δe_{new}，センサノードの保持情報量とシンクノードの距離の平均を d_{old}，1 ステップ後のそれを d_{new} として，報酬関数を

$$\frac{\Delta e_{old} - \Delta e_{new}}{\Delta e_{old}} + \left(\frac{d_{old} - d_{new}}{d_{old}}\right)^3 \qquad (10)$$

と設定する．

式 (10) の第 1 項は，観測している範囲内でセンサノード間のエネルギー残量差が小さくなるように作用し，第 2 項はセンサノードが持っている情報がシンクノードに近づくように作用する．

ここで式 (3) における $\phi(s, a)$ は，線形アーキテクチャ [11] を用いて求めることとする．データを送信したセンサノードのデータ送信前エネルギー残量，データ送信後エネルギー残量，データ送信前のセンサノードとシンクノードの距離，データ送信後にデータを保持しているノードとシンクノードの距離の 4 つの指標の特徴をそれぞれ三角分布の重み付き和を用いて近似することで $\phi(s, a)$ を決定する．

一例として，データ送信前エネルギー残量 e_b を特徴数 $n = 3$ として近似するときの線形アーキテクチャを**図 2** で表す．図中の w_1, w_2, w_3 は重みであり，これが e_b の線形アーキテクチャ $\phi_{e_b}(e_b)$ の形を決定する．同様に，他の 3 つの指標の特徴も三角分布の重み付き和で表すことができる．最後に，4 指標の特徴をベクトルに並べることで，$\phi(s, a)$ が得られる．

式 (3) で示すように確率的方策 π_θ は θ と $\phi(s, a)$ の線形結合で表され，$\phi(s, a)$ の各要素は重みベクトルと三角分布の値の線形結合で表されるため，求めたいパラメータ θ が重みベクトルを並べたものであると解釈できる．

3.4 部分情報に基づくマルチエージェント方式

マルチエージェント方式の場合においては，あるセンサノードと最近接最大 3 ノードのエネルギー残量平均の差を Δe_{old}，1 ステップ後の同じものを Δe_{new}，センサノードの保持情報量とシンクノードの距離を d_{old}，1 ステップ後のそれを d_{new} と捉えなおして，式 (10) により報酬関数を定義することで，あとは前節と同様に強化学習を行うことが可能となる特に，各センサノードが観測できる環境に共通部分があることから，全体の状態に関する知識がなくても，センサネットワーク全体のライフタイムが長くなるようなデータ通信経路が学習されると期待される．

4. シミュレーション評価

常に固定化された経路に従ってデータ通信をする場合と比較しつつ，強化学習を用いた自律的な通信経路決定手法の評価を行う．

4.1 シミュレーション条件

強化学習における各パラメータは，以下のように設定する．

- $M = 20,\ T = 15$
- 時間ステップ 700 回
- 学習率 $\eta = 10$

また，各センサノードの初期エネルギー残量，保持データ量，データ収集間隔は以下とした．

- $E_i = 5000$
- $I_i = \begin{cases} 1 & i = 9 \\ 0 & \text{otherwise} \end{cases}$
- $r_i = \begin{cases} 3 & i = 3, 5, 6 \\ 5 & \text{otherwise} \end{cases}$

データ通信コストは $E_t = 3,\ E_r = 1$ としてシミュレーションを行う．

強化学習を行わない場合には，データ通信経路として各センサノードがシンクノードに最も近づく最近接ノードにデータ通信を行う経路選択を考える．例えば，センサ $S7$ からは $S7 \to S4 \to S1 \to$ sink，$S8$ からは $S8 \to S5 \to S2 \to$sink，$S9$ からは $S9 \to S6 \to S3 \to$sink，というデータ通信経路に従い，各ノードがデータ送信を行う．

このように単純な通信経路に従って転送を行う場合，シンクノードに近いセンサノードを必ず通ってデータがシンクノードに集められるため，それらのノードのエネルギー残量の減りが速く，早い段階でバッテリ切れを起こすノードが出現する，つまりセンサネットワークのライフタイムが短くなると予想される．また，データ収集間隔の違いによってもライフタイムに影響すると想定される．

4.2 評価結果

本評価では，提案する強化学習を用いたデータ通信経路自律制御により，状況に適したデータ通信が行えていることを各センサノードのエネルギー残量を観測することで確認する．

図 3，**図 4**，および**図 5** に，強化学習を用いずにシンクノードに最も近づく最近接ノードに固定的に通信した場

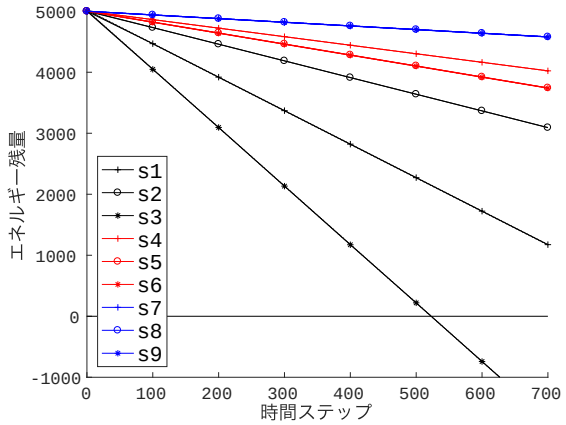

図 3 固定通信経路利用時のエネルギー残量変化

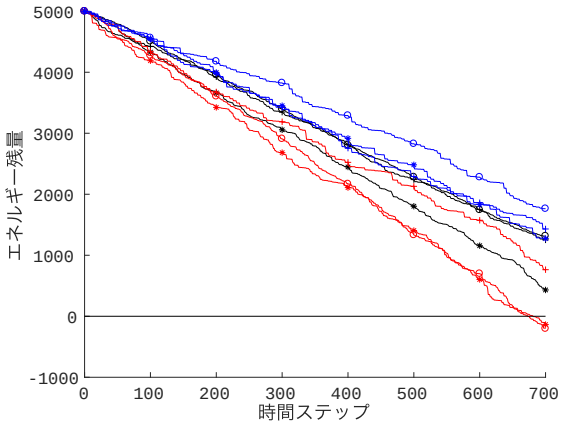

図 4 全体管理マスタ計算機方式の強化学習におけるエネルギー残量変化

合，マスタ計算機を利用した強化学習を行った場合，部分的に環境を観測できるマルチエージェント強化学習を行った場合の各センサノードのエネルギー残量の変化をそれぞれ示す．時間ステップが進むにつれて，各センサノードのエネルギー残量が減少し，またその減り方はセンサノードによって異なることがわかる．3 つのグラフを比較すると，強化学習を用いた提案手法では，強化学習を用いずに固定的な通信経路を利用する場合に比べて，エネルギー残量が 0 になるノードが出現する時刻が長くなっている，すなわちセンサネットワーク全体のライフタイムを長くできていることが確認できる．

表 1 は，それぞれの評価対象における，センサネットワークのライフタイム（単位は時間ステップ），初期状態から 500 時間ステップ後のエネルギー残量の平均と標準偏差をまとめたものである．なお，強化学習による行動選択は確率的なものであるため，10 回独立に強化学習を行いその平均値を求めることで各指標の値とした．この表より，強化学習を用いることでセンサネットワークのライフタイムが長くなる一方，500 時間ステップ後のエネルギー残量の平均は，強化学習を用いない場合が最も大きいことがわかる．エネルギー残量の標準偏差はマスタ計算機方式の強化学習手法が最も小さく，強化学習を行わない場合は最も大きい値となった．

4.3 考察

表 1 より，強化学習を用いない固定的通信経路手法が，エネルギー残量平均は最も大きいが，標準偏差も値が大きくなっている．エネルギー残量平均が大きい理由は，各センサノードが最近接のセンサノードまたはシンクノードにデータを通信しているため，全体の総消費エネルギーは最も小さくなるためである．この点では，通信のエネルギー効率が良いとも考えられるが，しかし，各センサノードがシンクノード側の最近接ノードにデータ通信を行うと，シンクノードに近いセンサノードは頻繁にデータ転送を行う必要があることから，シンクノードに近いノードと遠いノード間でエネルギー消費が大きく異なり，エネルギー残量の標準偏差が大きくなってしまう．さらに，この場合シンクノードに近いノードでは早い段階でバッテリ切れが発生しやすい．実際，図 3 では，$S1, S2, S3$ のエネルギー消費が大きく，$S3$ が最も早くバッテリ切れを起こしている．

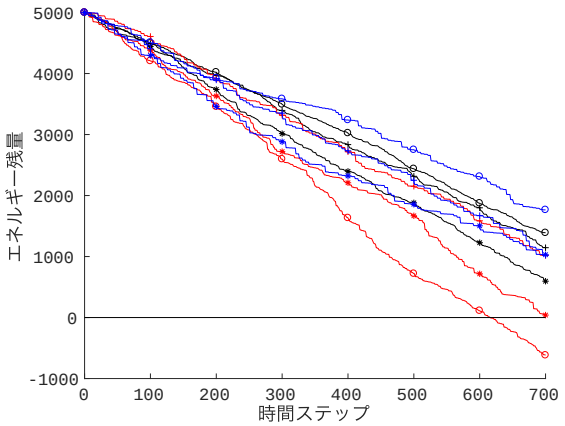

図 5 部分情報マルチエージェント強化学習におけるエネルギー残量変化

強化学習の有無によるエネルギー残量変化を比較すると，強化学習を用い自律的に制御した場合に，センサノード間のエネルギー残量差が小さくなるように，すなわち目的通りにデータ通信経路の制御が行えていることが分かる．これは式 (10) の報酬関数の第 1 項の効果である．また，センサネットワーク全体のライフタイムが長くなっていることから，式 (10) の第 2 項の効果として，エネルギーを無駄に消費した通信が抑制されていることが推察される．

センサノード間のエネルギー残量差が小さくなるように報酬関数を設計することで，シンクノードから遠いセンサノード群は，他ノードのエネルギー残量を考慮しつつデータ通信先を選択するため，エネルギー消費が固定経路手法に比べて大きくなる．その反面，シンクノードに近いセン

表 1　各手法の比較

	ライフタイム [時間ステップ]	エネルギー 残量平均	エネルギー 残量標準偏差
強化学習なし	523	3637	1491
強化学習 (全体)	627	2040	538
強化学習 (部分)	612	2035	561

サノード群は受信データ量が減り，エネルギー消費が抑制される．これにより，センサネットワーク全体のライフタイムを長くすることに貢献できており，強化学習を用いたデータ通信経路自律制御の有用性が確かめられたと考えられる．

図 4 と図 5 を比較すると，全体情報を用いたマスタ計算機方式のほうが，より多い観測情報で学習を行えているため，より良い方策が得られていることが分かる．しかし，前述の通り，全ての情報が観測でき，かつそれらが随時更新できるセンサーネットワーク環境は非現実的である．部分情報の観測のみで制御を行うマルチエージェント方式は，マスタ計算機方式に比べて多少劣ってはいるものの，十分有用な結果を得られている．これは，各センサノードが部分的な観測情報に基づき，自身の周囲にあるセンサノードとのエネルギー残量差を小さくする働きをするが，部分的に共有情報を持つためセンサネットワーク全体に効果が波及しているものと考えられる．

さらに，図 4 と図 5 から，提案手法ではデータ収集間隔が短いセンサノードのエネルギーの減り方も速くなっていることが確認できる．データの通信先を選択する際に，本稿ではエネルギー残量と距離の寄与を同等に扱うよう報酬関数を設計したが，それぞれの寄与の割合を変化させることで，よりセンサノードのエネルギー残量変化のばらつきを小さくし，センサネットワーク全体のライフタイムを更に長くすることも可能であると考えられる．

5. まとめと今後の課題

本稿では，マルチホップ WSN において強化学習を用いた自律的なデータ通信経路制御を実現するための方式を提案した．全体の情報を観測できるマスタ計算機を仮定した問題検討と，より現実的なモデルとしてマルチエージェント方式への拡張を検討し，報酬関数の設計を通して強化学習モデルを提案した．シミュレーション評価により，強化学習を用いた提案手法を用いることで，センサネットワーク全体のライフタイムを長くすることが可能であることを確認した．

今後は，マルチエージェント・センサネットワークモデルを改善することや，センサノードの配置が格子状ではなく，対称性のない場合に対しても提案手法の有用性を示すこと，Actor-Critic アルゴリズムや自然方策勾配法などのより正確な近似から導かれる強化学習アルゴリズムを用いることで，データ通信経路自律制御のより良い方策を得る方法などを検討していく予定である．

謝辞　本研究の一部は，JST CREST 課題番号 JPMJCR1785（研究課題名「リアルタイム性と全データ性を両立するエッジ学習基盤」）の支援を受けたものである．

参考文献

[1] K. Abe, T. Mizuno, H. Mineno: ”Development of the low consumption electricity type wireless node for wireless sensor networks,” IPSJ SIG Technical Report, Vol.2011-MBL-59, No.5, 2011
[2] H. -P. Shiang, M. van der Schaar: ”Online Learning in Autonomic Multi-Hop Wireless Networks for Transmitting Mission-Critical Applications,” IEEE Journal on Selected Areas in Communications, Vol. 28, No. 5, pp. 728-741, 2010
[3] X. Wu, G. Chen, S. K. Das, ”Avoiding Energy Holes in Wireless Sensor Networks with Nonuniform Node Distribution,” IEEE Transactions on Parallel and Distributed Systems, Vol. 19, No. 5, 2008
[4] G. Chen, C. Li, M. Ye, J. Wu, ”An unequal cluster-based routing protocol in wireless sensor networks,” Wireless Networks, Vol. 15, No. 2, pp. 193-207, 2009
[5] S. Yamawake *et al.*: ”Swarm Reinforcement Learning Method for Multi-agent Tasks,” T. SICE, Vol. 49, No. 3, pp. 370-377, 2013
[6] R. S. Sutton *et al.*: ”Policy Gradient Methods for Reinforcement Learning with Function Approximation,” Advances in Neural Information Processing Systems, Vol. 12, pp. 1057-1063, 2000
[7] R. J. Williams: ”Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning,” Machine Learning, Vol. 8, Issue 3, pp. 229-256, 1992
[8] S. Amari: ”Natural Gradient Works Efficiently in Learning,” Neural Computation, Vol. 10, No. 2, pp. 251-276, 1998
[9] S. Kakade: ”A Natural Policy Gradient,” Neural Information Processing Systems 14, pp. 1531-1538, 2001
[10] J. Peters, S. Schaal: ”Natural Actor-Critic,” Neurocomputing, Vol. 71, Issues 7-9, pp. 1180-1190, 2008
[11] I. Menache, S. Mannor, N. Shimkin: ”Basis Function Adaptation in Temporal Difference Reinforcement Learning,” Annals of Operations Research, Vol. 134, No. 1, 215-238, 2005