Vol.2017-ARC-225 No.26
Vol.2017-SLDM-179 No.26
Vol.2017-EMB-44 No.26
2017/3/9

強化学習を用いた環境発電駆動センサノードの適応的電力制御手法の検討

シュレスタマリ サソット[1,a)]　近藤 正章[1]　中村 宏[1]

概要：太陽光発電などの環境発電により駆動するセンサノードでは，バッテリ切れによるノードダウンを防ぎつつ，発電電力に応じてセンシング間隔を調整するなどの電力制御を行う必要がある．本稿では，機械学習の一手法である強化学習を用い，発電電力量や天気予報の情報をもとに センサノードのセンシング間隔を適応的に制御する手法を提案し，気象データを用いた評価を行う．評価の結果，提案手法は設置場所の変化などの運用環境の変化にも自動的に順応し高い性能を達成できることがわかった．

1. Introduction

膨大な数のセンサノードがネットワークに接続され，様々な情報を利用して我々の生活を豊かにする Internet of Things (IoT) 時代の到来が近づいている．そのうようなセンサノードは，一度設置されると，人手による保守なく高い信頼性を持ち，持続的に動作可能であることが望ましい．特に無線で通信を行うようなワイヤレスセンサノードにとって，エネルギーの供給をどうするかは重要な課題である．バッテリ駆動のシステムでは，省電力化によりバッテリ動作の寿命を長くさせることが可能ではあるが，それでも稼働可能時間には限りがある．そのため，バッテリの交換や再充電が必要となり，保守のコストが高くなる．センサノード自身のコストよりも、保守のコストの方がしばしば高くなるとさえ言われている [13].

そこで，この問題を解決するために，環境発電を利用してセンサノードを動作させる環境発電駆動センサーが注目されている．環境からエネルギーを得ることで，センサノードの設置場所の制約によらず，持続的なセンサの運用が期待できる [7]．この環境発電駆動センサノードでは，エネルギー消費を最小化することではなく，発電電力を利用していかに高い性能を発揮させるかに目的が変化する．そこで，エネルギー中立性 (energy neutrality)[7] が重要な概念となる．エネルギー中立性は，最大限にノードの性能を引き出しつつ，エネルギー消費と発電されたエネルギーが等しい状態を意味する．これらは，ENO-Max 条件 [17]，あるいはノードレベルエネルギー中立 (node level energy neutrality)[14] とも呼ばれる．

[1] 東京大学 大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo
[a)] shaswot@hal.ipc.i.u-tokyo.ac.jp

一方で，発電電力や電力消費，バッテリ容量などには限界がある．また環境発電においては発電電力は時間により変動し，時としてその予測が難しいなど，安定した電力供給が得られるわけではないなどの課題がある．そのため，様々な状況で ENO-Max 条件を達成することは簡単ではない．本稿では，ENO-Max 条件を達成するために、センサノードのデューティサイクルを調整する問題へと帰着させてこれを議論する．

デューティサイクルの調整に関しては，これまでにも多くの研究が行われてきた．例えば，発電されるエネルギー量を予測してデューティサイクルを調整しエネルギー中立性の達成を試みるものや [7]，線形 2 次トラッカあを利用するもの [17] なども提案されている．しかしながら，これらの手法では，設置された環境やバッテリサイズ，プロセッサの消費電力など，システムの仕様に応じていくつかのパラメータを調整しつつ運用することが必要となる．1 兆個にも及ぶセンサが様々な環境下で利用されると予想される将来の IoT 社会では，全てのセンサに対してパラメータを調整することは現実的ではない．

本稿では ENO-Max 条件を達成するためのより汎用的な手法として，強化学習を用いた手法を提案する．強化学習は，学習器が様々な一連の行動によりもたらされる結果を探索し，将来にわたり得られる報酬の合計が最大となるような行動を学習フェーズで記録する．そして，実行段階では，学習結果にもとづき得られる報酬が最大となる行動を状況に応じて選択するものである．本稿では，強化学習を用いたデューティサイクル調整手法を検討する．また，天気予報情報を利用して，センサノードの性能をより向上させる手法も提案する．なお，本稿では太陽光発電を利用したセンサノードを仮定して検討を行うが，提案手法は他の環境発電にも適用可能である．提案手法により手動での最

Vol.2017-ARC-225 No.26
Vol.2017-SLDM-179 No.26
Vol.2017-EMB-44 No.26
2017/3/9

適なパラメータ設定の手間を最小化しつつ，環境から得られるデータを基に ENO-Max 条件を達成し，また環境の違いにも適応的に対応できることが期待される．

2. 関連研究

環境発電の様々なアーキテクチャやそのエネルギー源，また実際の環境発電駆動センサノードの例が文献 [14] にまとめられている．また，文献 [9] では，無線センサノードの様々な電力管理戦略について述べられており，特にエネルギー供給と消費の面から無線センサノードの分類が行われている．

環境発電駆動センサノードについて最初に形式的に述べられているのが文献 [7] である．電力管理の基本的なアプローチとして，将来のタイムインターバルにおける発電エネルギー量を予測し，その情報をもとにデューティサイクルを決定する手法がとられている．発電エネルギー量の予測と実際の発電量が異なる場合への適応化について考慮する手法もある [3]．文献 [17] では，バッテリを中心に据えた目的関数を考案してこの問題へ対処している．

強化学習を用い，バッテリレベルと環境発電レベル，エネルギー中立性を考慮したデューティサイクル最適化手法が文献 [4] で提案されている．我々の手法は，これらの手法をベースに，環境への適応性やパラメータ調整法を向上させ，より良い性能が得られるようにしたものである．また，天気予報情報を利用して将来の発電エネルギー量を考慮する点も本稿の提案手法の新規性である．文献 [4] の著者らは，その手法を拡張し，スループット要求を意識したものも提案している [5]．また，ファジー論理を用いて状態や報酬の見積もりを行う拡張手法も提案されている [6]．

連続時間マルコフ連鎖モデルを用い，バッテリ特性で異なるバッテリ充電率や QoS 要求を考慮し，適応的にデューティサイクルを調整する手法が文献 [2] で述べられている．文献 [1] では，ポイント・ツー・ポイントの無線通信システムにおける確率的なデータ到着モデルやチャネル状態を意識した手法が提案されている．有限のバッファサイズのもとで，合計送信データ量を最大化させるための手法を検討するものもある [10]．報酬や状態が不確かなシステムにおけるベイジアン強化学習を用いた最適化手法の検討が文献 [18] で行われている．また，文献 [12] では，Q 学習ににおける次元の呪いの問題へ対処するために，関数近似を用い，エネルギー中立性を失うことなく，スループットを最大化させる手法を提案している．

これらの従来手法は，センサノードにおけるバッテリの劣化や設置場所の変化など，運用環境のパラメータが変わった場合にどのように振る舞いを順応させるかについては述べられてはいない．本稿では，より汎用的に用いることのできる手法として，運用環境が学習時とは異なる場合にも人手を介することなく環境に適応し，高い性能を得ることができる手法を提案する点で，従来研究とは異なる．

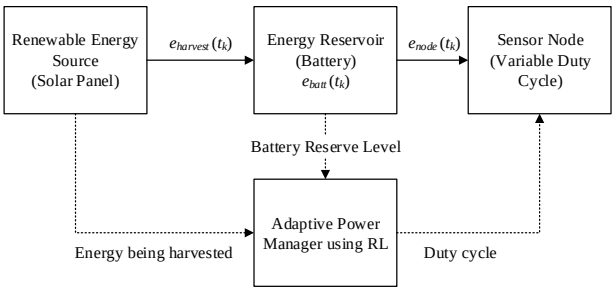

図 1 システムモデル

3. 理論

3.1 システムモデル

ここでは，検討対象の環境発電駆動センサノードのシステムモデルとして，環境発電エネルギー源，バッテリ，エネルギーを消費する側のセンサノード，そして電力管理ユニットの主として 4 つの構成部からなるモデルを検討する (図 1 参照)．センサノードのデューティサイクルは可変であり，デューティサイクルを上げることで性能が高くなるとする．時間はエポックと呼ぶ等区間に分割して考える離散時間モデルを仮定する．各エポック t_k では，システムは環境から有限のエネルギー $e_{harvest}(t_k)$ を取得する．電力管理ユニットはシステムのデューティサイクル $d(t_k)$ を決定する．センサノードは $d(t_k)$ に応じてエネルギー $e_{node}(t_k)$ を消費する．システムには容量 B_{full} のエネルギーを蓄えることのできるバッテリが接続される．$e_{batt}(t_k)$ は各エポック t_k の開始時点でのバッテリ残量とする．エポック t_{k+1} でのバッテリ残量は以下の式で表される:

$$\Delta e_{batt}(t_k+1) = e_{batt}(t_k) + e_{harvest}(t_k) - e_{node}(t_k) \quad (1)$$

エネルギー中立状態は，発電エネルギーとエネルギー消費とが均衡している状態であり，あるエポック t_k におけるエネルギー中立状態からの差分 $\Delta e_{neutral}(t_k)$ は以下のようになる:

$$\Delta e_{neutral}(t_k) = e_{harvest}(t_k) - e_{node}(t_k) \quad (2)$$

ある時間間隔で，バッテリが完全に充電された後にも余剰のエネルギーがあった場合，そのエネルギーは「無駄になった」ことになる．ここで，$[x^+]$ を以下のように定義する:

$$[x^+] = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (3)$$

これを用い，エポック t_k において無駄になったエネルギーは以下のように計算される:

$$e_{waste}(t_k) = [e_{batt}(t_k - 1) + e_{harvest}(t_k) - e_{node}(t_k) - B_{full}]^+ \quad (4)$$

3.2 強化学習

強化学習 (Reinforcement Learning: RL) は，教授するのではなく経験から学ぶことをベースとした機械学習の一手法である．強化学習により，システムが種々の行動から最適と考えられる行動を自動で選択することが可能となる．好ましい行動をすると報酬が与えられ，好ましくない行動の場合には罰が与えられる．試行錯誤を繰り返すことで，学習により長期的な報酬を最大化するような行動をシステムが選択することが可能となる．

一般的な強化学習モデルでは，有限の状態空間 S，行動集合 A の中の行動を実行可能なエージェント，行動に反応して状態が変化する環境から成る．ある状態で行動を行うことより，エージェントは報酬関数 $R : S \times A \to R$ にもとづいてスカラ量の報酬を受け取る．エージェントがある状態 s にいる際に選択する行動 a は，方策 $\pi = \{(s, a)|a \in A, s \in S\}$ にもとづいて決定され，$\pi(s) = a$ として表される．

エージェントと環境との相互作用はイベントと呼ぶ離散時間のステップで行われ，各イベントは一つのエポック内で発生する．あるエポック t_k ではエージェントは状態 $s_k \in S$ を持ち，ある方策 π にもとづいて行動 $a_k \in A$ を行う．環境はとられた行動に反応して，エージェントの状態を次の状態 $s_{k+1} \in S$ に変化させるとともに，報酬 r_k を与えることになる．

エージェントの目的は，将来にわたり得られると期待される (割引を含む) 報酬の合計が最大となるような各時間ステップでの方策を見つけることである．全ての状態と行動のペアについて最良の方策 π^* がわかれば，報酬の合計は最大化される．

Q 学習 (Q-learning) は強化学習のアルゴリズムの一つであり，各状態–行動のペアに対して Q 値を割り当て，それを Q テーブルに保存する．ある状態–行動のペア (s, a) の Q 値は，状態 s を開始点として，最初の行動として行動 a を選択した場合に得られる割引を加味した報酬の和の最大値と定義される [15]．言い換えると，ある状態–行動のペア (s, a) の Q 値は，状態 s から開始し，行動 a を選択し，その後は最良の方策 π^* にしたがって行動した場合の合計の報酬の期待値となる．これは以下の式で表される:

$$Q(s, a) = E\left[\sum_{k=0}^{N-1} \gamma^k r(s_k, a_k)\right] \tag{5}$$

ここで，$s_0 = s, a_0 = a$，また $a_k = \pi^*(s_k)$ である．

Q 学習は各状態–行動ペアの Q 値を決定するプロセスである．一度 Q 値が確定すれば，その後は最良の行動を決定することは簡単である．各状態において，最も高い Q 値を持つ状態へと移るような行動を選択すれば良い．

Q 学習においては，エポック t_k から t_{k+1} へと移る時に，学習器は s_k, a_k, r_k と s_{k+1} を観測する．これらの値をもとにエポック t_{k+1} で Q 値の推定値を更新する [16]．言い換えると，学習中には，ある推定値を他の部分的な推定値をもとに更新していくことを続けることになる．ある状態–行動ペアの Q 値の更新は以下の式にもとづいて行われる．

$$Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha\left[r_k + \gamma \max_a, Q(s_{k+1}, a) - Q(s_k, a_k)\right] \tag{6}$$

ここで，定数 α は各ステップで Q 値をどれだけ変更するかを決定するための学習率を表す．全ての状態から全ての行動を選択することを続けていくことで，Q 値の推定値が最終的に真の Q 値に収束することが証明されている [11], [17].

3.2.1 Greedy 法

Greedy 法では，エージェントは現在の Q テーブルの情報をもとに，常に Q 値が最大の行動を選択する．そのため，Q 値が最適な値に収束していない場合には，局所最適な行動のみを選択してしまう．

3.2.2 ϵ-greedy 法

大規模な Q テーブルの場合，多数の試行錯誤を繰り返したとしても収束させることは難しい場合も存在する．そこで，ϵ-greedy 法では，エージェントは一定の確率 ϵ でランダムに行動を選択し，他の行動の価値を探索する．この手法の利点は，経験を積む毎に，より多くの行動が複数回にわたって評価されることになり，局所最適な状態になることを防げることである．

4. 環境発電駆動センサノード向け強化学習モデル

ここでは，提案する強化学習モデルと Q 学習のための問題設定について述べる．

4.1 Q 学習の問題設定

まず，強化学習モデルで重要となる状態の集合と行動の集合について定義する．本モデルでは，環境はバッテリと確率的に振る舞うエネルギー源とから構成される．環境はエージェントに対してデューティサイクルに応じた報酬を与え，発電エネルギー量とバッテリ残量レベルに対応して新たな状態を指定する．状態の定義，および行動の集合と報酬の仕組みについて以下に詳述する．

4.1.1 状態

状態の集合 S は，バッテリ残量レベルの状態 S_{batt}，一つ前のエポックにおける発電エネルギー量の状態 $S_{eharvest}$ からなる．そのため，状態空間は S_{batt} と $S_{eharvest}$ の組み合わせ，すなわち $(S_{batt}, S_{eharvest}) \in S$ で与えられる．

4.1.2 行動の集合

行動の集合 A は設定可能なデューティサイクルの集合 $A \in \{D_{min}, D_{max}\}$ で定義される．ここで，D_{min} と D_{max} はそれぞれ，対象とするセンサノードで設定可能なデューティサイクルの最小，および最大値である．エージェントは各エポックで 1 つのデューティサイクルのみ設定することができる．

4.1.3 報酬関数

報酬関数とは，エージェントにとってどのような振る舞いが好ましいか好ましくないかを指定するものである．こ

こでは，エポック t_k におけるバッテリ残量レベル，発電エネルギー量，ノードでのエネルギー消費量の 3 つの要因に依存して報酬が決まるような報酬関数 r_k を定義する．具体的には，以下のように報酬関数を定める:

$$r_k = \frac{B_{th} - e_{batt}(t_k)}{B_{full}} \times \frac{e_{harvest}(t_k) - e_{node}(t_k)}{e_{maxh} - e_{minh}} \quad (7)$$

ここで，B_{th} はバッテリ残量に関する閾値であり，e_{maxh} と e_{minh} はそれぞれ最大，最小の発電エネルギー量である．

また，バッテリ残量が極端に低下すると動作を続けられなくなるため，エージェントがそのような状態になるべく陥らないよう，以下に述べる 2 つの条件付き報酬をさらに加えることにする．もし，$e_{batt}(t_k)$ が下の閾値レベル B_{thl} よりも低下した場合は，報酬は 25%削減される．さらに，$e_{batt}(t_k)$ が下限の閾値レベル B_{thll} よりも低下した場合は報酬は 75%削減される．これらの閾値レベルはセンサノードが置かれる環境に依存して，ユーザにより設定される．

4.1.4 学習アルゴリズム

本稿では，エージェントにおける状態–行動ペアの Q 値を決定するために，先に述べた Q 学習を用いる．収束を保証させるために，学習が進むにつれて学習パラメータである α の値を小さくなるようにする．学習フェーズでは，全パターンの状態遷移が発生し，該当する Q 値が推定できるように探索的に行動を選択する．また，過去の気象データを用い十分に Q 値の値が収束するまで学習を行う．

4.2 実装

Q テーブルが十分に収束したところでセンサノードを配置すると仮定し，あとは Greedy 法にしたがって行動を選択するポリシーを用いる．もし，センサノードの運用中に環境の状況が大きく変化する場合は，ϵ−greedy 法を用いる．

5. 評価

5.1 システムの仮定

- **エネルギー源:** 本稿では，太陽光パネルを持つセンサノードを想定し，気象庁のウェブサイト (*http://www.jma.go.jp)* より全天日射量デー タを取得して，発電エネルギー量を計算する．当該データには，国内のいくつかの場所について，時間当たりの全天日射量が保存されている．本稿では，このデータに合わせ，エポックは *1* 時間として評価を行う．
- **センサノード:** センサノードは指定したデューティサイクルに応じてエポックあたり *50mWh* から *500mWh* のエネルギーを消費すると仮定する．また，エポック内では消費電力に変動はなく，一定の消費電力であると仮定する．デューティサイクルは *10%, 20% ...100%*のように，*10%*刻みで設定可能とした．デューティサイクル変更にともなうレイテンシなどのオーバーヘッドは無視できるものとする．
- **バッテリ:** バッテリとしては，容量が *20000mWh* で理想的な特性を持つバッテリを仮定し，*round-trip* 効果や漏れ電流，時間経過にともなうバッテリの劣化は考慮しない．充電ロスや，漏れ電流はノードにおける余分な消費エネルギーと見なすことが可能である．

5.2 Q 学習の仮定

- **状態**
  システムの状態は $S \in \{S_{batt}, S_{eharvest}\}$ で定義され，それぞれ $S_{batt} \in \{1, 2...200\}$，および $S_{eharvest} \in \{1, 2...5\}$ とする．前者は，バッテリのフル充電状態 B_{full} を 200 等分したものであり，後者は発電エネルギー量を 500mWhr 単位で表したものとなる．さらに，より高度な制御のため，将来の気象データとして天気予報を含めた状態の定義をする．天気予報データを含める場合は，状態は $S \in \{S_{batt}, S_{eharvest}, S_{fcast}\}$ で表され，S_{fcast} は翌日の総発電エネルギー量として，$S_{fcast} \in \{1, 2...8\}$ と区分して定義する．
- **Action Space**
  行動の集合 A は各エポックで選択するデューティサイクルで定義される．評価では $A = a(k) \in \{10\%, 20\%, 30\%...100\%\}$ とし，$a(k)$ はエポック t_k のパーセンテージ表記のデューティサイクルである．
- **Reward Function** 報酬関数については，バッテリの閾値を B_t =6000mWh，B_{thl} =4000mWh，B_{thll} =2000mWh とする．e_{maxn} は 3000mWh と仮定し，e_{minh} は 60mWh とする．

6. 評価結果

6.1 基本評価

まず，提案する強化学習を用いた手法と，他の 2 つの電力管理手法とで比較評価を行う．1 つ目は Naive な手法であり，バッテリ残容量が下限の閾値を下回った場合は最も低いデューティサイクルへ，上限の閾値を上回った場合は最も高いデューティサイクルへ，その中間であればバッテリ残容量に比例する値を利用するというように，バッテリの残容量に応じてデューティサイクルを決定するものである．2 つ目は Kansal らによる従来研究 [7], [8] をベースとしたもので，1 日の合計発電エネルギー量を予測し，その情報をもとにエネルギー中立性が確保されるよう一日を通して一定のデューティサイクルを決定する．なお，1 日の発電エネルギー量は完全に予測できるものと仮定する．

また，提案手法においても将来の気象情報として天気予報を利用するかどうかが異なる 2 種類を評価する．将来の気象情報を用いない場合 (RL) はバッテリレベルと発電エネルギーを状態として，また用いる場合 (RL with forcast) はさらに将来の発電エネルギー量を状態に加えたものとなる．この際に，将来の発電エネルギー量は常に予報が当たった場合として正しい情報を利用をする．両者ともに，

Vol.2017-ARC-225 No.26
Vol.2017-SLDM-179 No.26
Vol.2017-EMB-44 No.26
2017/3/9

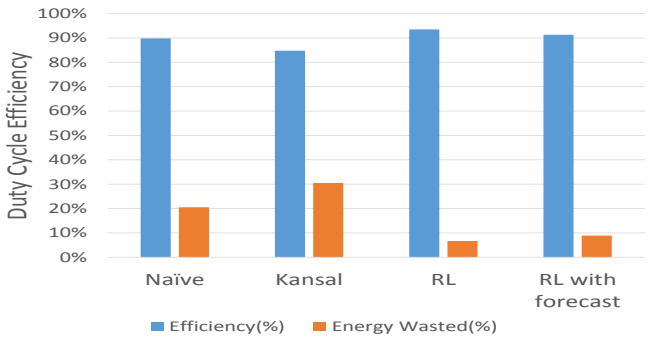

図 2 提案手法と従来手法の比較

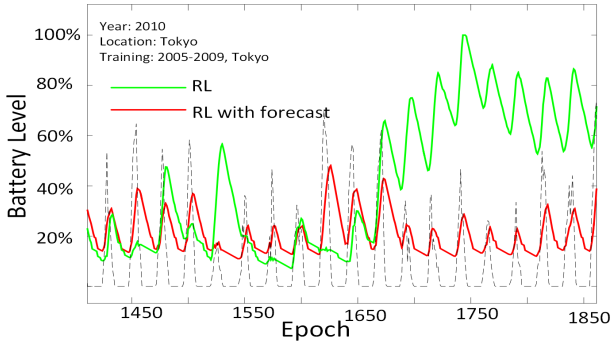

図 3 バッテリ残量レベルの時系列変化

東京地点の 2000 年から 2009 年の日射量データを用いて学習し，後は greedy 法に従うものとする．

図 2 に，2010 年の東京地点のデータを利用して評価を行った結果を示す．評価指標としては，理想的に最大で達成可能な平均デューティサイクルに対する実際に選択された平均デューティサイクル (*Efficiency*) と，年間を通した全発電エネルギー量に対してバッテリフルにより使われなかったエネルギー量 (*Energy Waisted*) を用いている．図より，提案手法の RL が他の従来手法よりも両評価指標において優れていることがわかる．

次に，天気予報情報の利用の効果について考察する．図 2 を見ると，天気予報情報を利用した場合にはかえって Efficiency が低下している．これは，天気予報情報を利用することで，状態の数が相当に増加し，Q 値の収束が困難になったためである．

一方で，天気予報情報を利用には利点もある．両者の振る舞いの違いを見るために，図 3 にエポック毎のバッテリ残量レベルを時系列で示す．1650 エポック付近より両者の手法でバッテリ残量レベルが大きく異なっているが，これは天気予報情報を利用する手法では，翌日の天気予報データを加味することで，より安全側に，すなわち低いデューティサイクルを利用するように制御が行われたからである．一方で，天気予報情報を利用しない場合は，積極的に高いデューティサイクルが利用され，バッテリ残量が危険域まで低下してしまっている．バッテリ切れは環境発電駆動センサノードにとって最も避けるべき事象であり，性能は多少低下するものの，天気予報情報を利用することでより安全に制御できる点は意義があると考えられる．

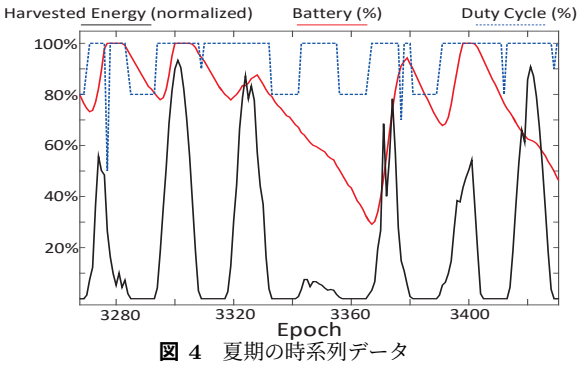

図 4 夏期の時系列データ

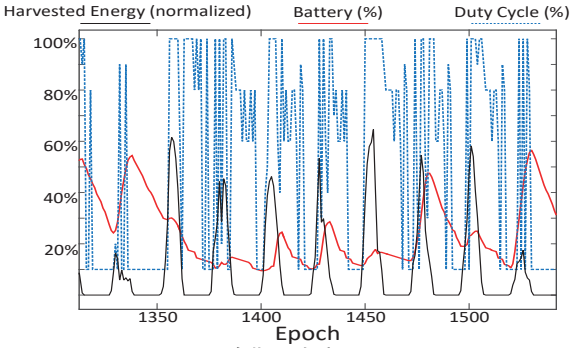

図 5 冬期の時系列データ

6.2 季節変化への適応性

ここでは，提案手法が季節によるエネルギー発電量の違いに適応できているかを考察する．前節と同じく，東京地点の 2000 年から 2009 年の日射量データにより学習し，2010 年のデータで評価を行った際の夏季の時系列データを図 4 に，冬季の時系列データを図 5 に示す．なお，天気予報情報は用いていない．図 4 より，夏季には日照時間も長く発電エネルギー量も多いため，提案手法では積極的に最大のデューティサイクルを利用している．また，例えば 3340〜3360 番目のエポックのように，発電がない夜間においても高いデューティサイクルが達成されていることもわかる．

対照的に，図 5 の冬季の結果では，提案手法ではそれほど高いデューティサイクルを用いようとはしていない．ただし，バッテリ残量の状況に応じてデューティサイクルは調整されており，時々高いデューティサイクルでの実行を試みていることがわかる．このように，季節に応じて提案手法では適切にデューティサイクルが制御されている．

6.3 場所の変化への適応性

発電エネルギー量は，センサノードが設置された場所により大きく異なる．学習時と異なる場所に設置された場合にも，発電エネルギー量の違いに適切に対応できることは重要である．そこで，本実験では，東京地点の 2000 年から 2009 年の日射量データにより学習を行い，気候条件が異なるであろう複数の地点でのデータを適用して評価を行う．

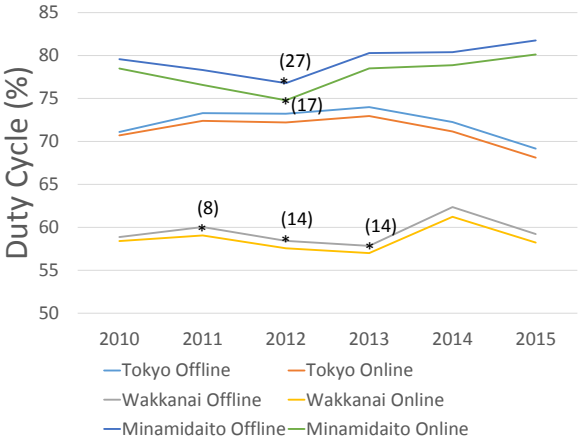

図 6 Greedy および ϵ-greedy 法による性能の違い

本評価では特に，東京以外に，南大東と稚内のデータを利用して評価を行う．この際に，Greedy 法と ϵ-greedy 法の比較も行う．ϵ-greedy 法はオンラインでの学習が可能である．図 6 に両手法を用いた場合の各年の平均デューティサイクルを示す．括弧内の数字は当該年にバッテリ切れが発生した回数を示している．

注目すべきは稚内での 2011 年から 2013 年の結果であり，Greedy 法ではバッテリ切れが数回発生しているにも関わらず，ϵ-greedy 法によりオンラインで学習を行うことでそれが回避されている．したがって，学習場所と設置場所が異なる場合にもオンライン学習を行う提案手法が適切に対応できていると言える．一方で，Greedy 法に比べて ϵ-greedy 法では，若干ではあるが，平均デューティサイクルが減少している．これは，ϵ-greedy 法では，探索を行うため，最適とは限らない行動をランダムに選択する必要があるためである．

7. おわりに

本稿では強化学習を用いた環境発電駆動センサノードのデューティサイクル管理手法を提案し，気象データを用いた評価を行った．評価結果では，提案手法により人手によるチューニングなしに環境からの状態のみを入力としているにも関わらず，ENO-Max 条件が達成されていることがわかった．また，季節の変化や，設置場所の違いにも適応可能であることを示した．これらの結果より，提案手法は様々な場所に膨大な数のセンサが設置されると IoT 時代の環境発電駆動センサノードの電力管理手法として有用なものになると結論付けることができる．天気予報を利用した際の Q 値の収束性を向上させるなど，さらに手法を改善して性能を向上させることなどが今後の課題である．

参考文献

[1] Blasco, P. et al.: A learning theoretic approach to energy harvesting communication system optimization, *IEEE Tr. on Wireless Communications*, Vol. 12, No. 4, pp. 1872–1882 (2013).
[2] Chan, W. H. R. et al.: Adaptive duty cycling in sensor networks with energy harvesting using continuous-time Markov chain and fluid models, *IEEE Journal on Selected Areas in Communications*, Vol. 33, No. 12, pp. 2687–2700 (2015).
[3] Hsu, J. et al.: Adaptive duty cycling for energy harvesting systems, *Proc. of the 2006 ISLPED*, pp. 180–185 (2006).
[4] Hsu, R. C. et al.: Reinforcement learning-based dynamic power management for energy harvesting wireless sensor network, *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, Springer, pp. 399–408 (2009).
[5] Hsu, R. C. et al.: A Reinforcement Learning-Based ToD Provisioning Dynamic Power Management for Sustainable Operation of Energy Harvesting Wireless Sensor Node, *IEEE Tr. on Emerging Topics in Computing*, Vol. 2, No. 2, pp. 181–191 (2014).
[6] Hsu, R. C. et al.: Dynamic energy management of energy harvesting wireless sensor nodes using fuzzy inference system with reinforcement learning, *IEEE 13th INDIN*, pp. 116–120 (2015).
[7] Kansal, A. et al.: Power management in energy harvesting sensor networks, *ACM Tr. on Embedded Computing Systems*, Vol. 6, No. 4, p. 32 (2007).
[8] Kansal, A., Potter, D. and Srivastava, M. B.: Performance aware tasking for environmentally powered sensor networks, *ACM SIGMETRICS Performance Evaluation Review*, Vol. 32, No. 1, pp. 223–234 (2004).
[9] Khan, J. A. et al.: Energy management in wireless sensor networks: a survey, *Computers & Electrical Engineering*, Vol. 41, pp. 159–176 (2015).
[10] Mao, S. et al.: Joint energy allocation for sensing and transmission in rechargeable wireless sensor networks, *IEEE Tr. on Vehicular Technology*, Vol. 63, No. 6, pp. 2862–2875 (2014).
[11] Moser, C. et al.: Adaptive power management in energy harvesting systems, *Proc. DATE*, pp. 773–778 (2007).
[12] Ortiz, A. et al.: Reinforcement learning for energy harvesting point-to-point communications, *2016 IEEE International Conference on Communications*, pp. 1–6 (2016).
[13] Raghunathan, V., Kansal, A., Hsu, J., Friedman, J. and Srivastava, M.: Design considerations for solar energy harvesting wireless embedded systems, *Proceedings of the 4th international symposium on Information processing in sensor networks*, IEEE Press, p. 64 (2005).
[14] Sudevalayam, S. and Kulkarni, P.: Energy harvesting sensor nodes: Survey and implications, *IEEE Communications Surveys & Tutorials*, Vol. 13, No. 3, pp. 443–461 (2011).
[15] Sutton, R. S. and Barto, A. G.: *Reinforcement learning: An introduction*, Vol. 1, No. 1, MIT press Cambridge (1998).
[16] Sutton, R. S. et al.: Reinforcement learning is direct adaptive optimal control, *IEEE Control Systems*, Vol. 12, No. 2, pp. 19–22 (1992).
[17] Vigorito, C. M. et al.: Adaptive control of duty cycling in energy-harvesting wireless sensor networks, *4th IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*, pp. 21–30 (2007).
[18] Xiao, Y. et al.: Bayesian reinforcement learning for energy harvesting communication systems with uncertainty, *2015 IEEE International Conference on Communications*, pp. 5398–5403 (2015).