Vol.2017-ARC-227 No.28
2017/7/27

適応的電力制御を行う環境発電駆動センサノードの強化学習戦略の比較評価

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，環境発電駆動センサノードとして，太陽光パネル，バッテリー，汎用のセンサノードデバイスからなる単純なシステムモデルを仮定し，機械学習の一手法である強化学習を用いた適応的な電力管理手法の比較評価を行う．強化学習を用いることで，天候やバッテリの劣化などの環境の変化に対して適応的に電力管理を行うことが可能になると期待される．比較に際しては，SARSA 学習と Q 学習アルゴリズムを用い，また適正度の履歴の有無による強化学習手法を評価する．評価の結果，SARSA(λ) 手法が他の手法に比べて優れた性能を達成できることがわかった．

1. はじめに

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) 時代が到来しつつある．バッテリ駆動のセンサノードが主流となると予想されるが，バッテリ交換などの保守コスト低減のために，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 [8]．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，エネルギー中立オペレーション (energy neutral operation) [8], [9], [15] が重要な概念となる．エネルギー中立オペレーションは，ノードの消費電力量が環境発電電力量と等しいかそれよりも少ない場合に達成される．さらに，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は ENO-Max 条件 [18], あるいはノードレベルエネルギー中立 (node level energy neutrality)[16] と呼ばれる．

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，様々な状況でノードレベルエネルギー中立を達成することは簡単ではない．

ノードレベルエネルギー中立達成において重要な点は，発電電力の変化に応じて適切な電力管理を行うことである．センサノードは，天候や発電電力効率の変化，バッテリーの劣化 [12]，さらにはノードの一部の故障などに応じて振る舞いを調整する必要がある．これら全事象や全状況に対して事前に対応策を準備することは現実的ではない．特に，個々のデバイスの電力消費傾向が異なる可能性のある 1 兆個にも及ぶセンサノードが，それぞれ特有の環境に配置されると考えられる IoT 時代ではなおさらである．そこで，電力管理戦略を学習し，環境へ配置後にも自律的に環境の変化に適応できるセンサノードを開発することが重要となる．

適応的な電力管理手法に関しては，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 [8]，線形 2 次トラッカを利用するもの [18] なども提案されている．電力管理に強化学習を用いる手法も提案されている [1], [4], [6], [14], [20]．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． 強化学習を用いたアプローチは自然と適応性を持つことができるが，学習には

[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

環境との相互作用が必要であり，刻々と変化する環境に対しての適応速度は比較的ゆるやかである [17]．そのため，強化学習による電力管理を考える上で，環境への適応性を評価・検討することは重要な課題である．

本稿では，2 つの異なる強化学習アルゴリズムである Q 学習と SARSA 学習を比較評価する．さらに，それぞれについて，学習に適正度の履歴 (eligibility trace) を用いる場合と用いない場合を検討する．適性度の履歴を利用することで，報酬の一部が結果に影響する全行動と状態に伝搬される．これにより，1 ステップ分の報酬を伝搬させる従来手法に比べて学習が高速に行える [17]．なお，本稿での問題設定では，センサノードはエネルギー中立オペレーション達成のために，デューティサイクルを調整することで消費電力を変更させると仮定する．電力管理機構は，エネルギー中立からの差分，バッテリー残量，環境発電電力量，その日の天気予報の情報をもとにデューティサイクルを決定する．ここで，エネルギー中立からの差分をエネルギー中立性能 (Energy Neutral Performance: ENP) と呼ぶことにする．

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 [16] にまとめられている．文献 [3] では予測可能，および不可能な環境下について，環境発電を用いた通信システムの課題と解決法について述べられている．また，文献 [10] 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 [8] 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 [4]．文献 [18] 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 [5] 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 [5] 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 [6]．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 [7]．

連続時間マルコフ連鎖モデルを用い，バッテリ特性で異

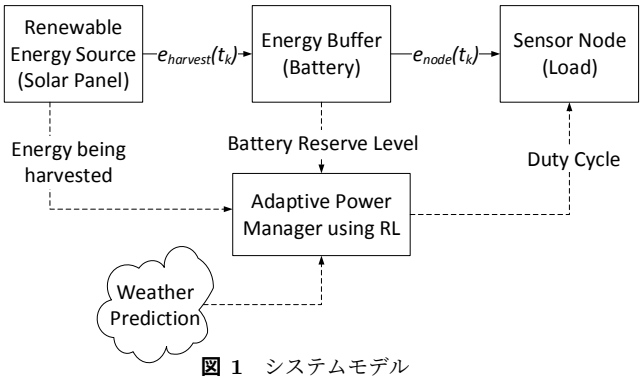

図 1 システムモデル

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 [11]．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 [19] で行われている．また，文献 [14] 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，種々の環境下での強化学習を用いたセンサノードの電力管理機構の振る舞いを調査し，適応性を評価する点がその貢献である．

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_{MAX} 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:

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

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 $e_{dist}(t_k)$ は以下のようになる:

$$e_{dist}(t_k) = e_{harvest}(t_k) - e_{node}(t_k) \qquad (2)$$

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．

エージェントが状態 s のとき，方策 π に従い行った行動 a 良さを評価するために，状態–行動のペアに対して Q 値 [17] を割り当てる．Q 値 $Q^\pi(s, a)$ は，状態 s から開始し，行動 a を選択し，その後は方策 π にしたがって行動した場合の割引を加味した合計の報酬の期待値と定義される．これは以下の式で表される:

$$Q^\pi(s, a) = E\left[\sum_{k=0}^{N-1} \gamma^k r(s_k, a_k)\right] \qquad (3)$$

ここで，$s_0 = s, a_0 = a$，また $a_k = \pi(s_k)$ である．γ ($0<\gamma<1$) は報酬の割引率である．なお，最良の方策 π^* の Q 値を Q^* と表す．本稿では，Q 値を学習するための 4 つの異なる手法として SARSA(λ)，SARSA，Q(λ) 学習，そして Q 学習を比較する．一度 Q^* 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り，各状態 s について，$Q^*(s, a)$ を最大化する行動 a を選択すれば良い．このように行動を選択する方法を greedy ポリシーと呼ぶ．

3.2.1 Q 学習

Q 学習においては，学習プロセスとして Q 値の推定値を更新していく．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 [17]．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．

$$Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha\left[r_k + \gamma \max_a Q(s_{k+1}, a) - Q(s_k, a_k)\right] \qquad (4)$$

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 [13], [18]．Algorithm 1 に Q 学習のプロセスを示す．

Algorithm 1 Q-learning algorithm

```
1: procedure Q–LEARNING
2:     Initialize Q(s, a) = qinit
3:     Observe current state s_k
4:     for all time do
5:         Choose and execute action a_k
6:         Receive reward r_k
7:         Observe new state s_{k+1}
8:         Update Q(s, a) according to Equation 4
9:         s_k = s_{k+1}
10:    end for
11: end procedure
```

3.2.2 SARSA

SARSA は State-Action-Reward-State-Action の頭文字をとったものである．ポリシー π に対応する状態–行動ペア (s_k, a_k) の Q 値 $Q^\pi(s, a)$ は，エージェントの他の状態–行動ペア (s_{k+1}, a_{k+1}) への遷移と受け取る報酬 r_k から推定される．Q 学習と同様に，エージェントは状態 s_k から開始し，方策 π にしたがって行動 a_k を実行する．結果として報酬 r_k を受け取り，他の状態 s_{k+1} に移行する．ここで，エージェントは方策 π にしたがって次の行動 a_{k+1} を実行することを検討する．この Q 値を更新する際に次の行動を検討する点が Q 学習と SARSA の重要な違いである．$Q(s_k, a_k)$ の更新は以下の式に基づいて行われる．

$$Q^\pi(s_k, a_k) \leftarrow (1-\alpha)Q^\pi(s_k, a_k) + \alpha\left[r_k + \gamma Q^\pi(s_{k+1}, a_{k+1})\right] \qquad (5)$$

SARSA 学習のアルゴリズムは，ステップ 8 で $Q(s, a)$ の更新に式 (5) を用いる点以外は Algorithm 1 と同一である．

3.2.3 適正度の履歴

エージェントがある連続して行動を実行し，一連の状態を経て報酬を受け取る際に，エピソードの最後にどのように信用割り当てを行うかが問題となる．この問題の解決のため，適正度の履歴 (*eligibility trace*) と呼ぶ変数を各状態–行動ペアに導入する．各エポック t_k の状態–行動ペアの適正度の履歴を $e_k(s,a) \in \mathbb{R}_{\geq 0}$ と表す．各エポックでは，全状態–行動ペアの適正度の履歴は $\gamma\lambda$ の割合で減衰させられ，エポック t_k で訪れた状態–行動ペアの履歴には 1 が加えられる．つまり，全 (s,a) に対し

$$e_k(s,a) = \begin{cases} \gamma\lambda e_{k-1}(s,a) & \text{if } (s,a) \neq (s_k,a_k) \\ \gamma\lambda e_{k-1}(s,a) + 1 & \text{if } (s,a) = (s_k,a_k) \end{cases} \quad (6)$$

となる．ここで，λ $(0 < \lambda < 1)$ は，最終的な報酬をもとに以前の状態–行動ペアの Q 値を更新する際の強度のパラメータである．$e(s,a)$ の値は，エピソードの最後に，状態–行動ペア (s,a) が得られた報酬に対してどの程度影響したかの指標となる．

3.2.4 ϵ-greedy ポリシー

Q テーブルは，Q 値の初期値として全状態–行動ペアに対して高い Q 値を与えることで楽観的に初期化される [17]. これは，学習開始当初から幅広く状態–行動ペアを探索するためである．十分に学習が行われ，全状態–行動ペアをくまなく訪れることで Q 値は収束する．しかし，greedy ポリシーに従うと，エージェントが良い結果をもたらすような状態–行動ペアを訪れることができないかもしれず，最適ではないポリシーが導かれる可能性がある．そこで，本稿では ϵ-greedy 法を利用する．この方法では，ほとんどの場合，エージェントは greedy ポリシーに従うが，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．

4. エネルギー中立向け強化学習手法

ここでは，本稿で扱う問題のための強化学習フレームワークと SARSA(λ) 学習アルゴリズムについて述べる．

4.1 強化学習フレームワーク

強化学習フレームワークとして，状態の集合と行動の集合について定義する．本フレームワークでは，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，天気予報情報に対応して新たな状態を指定する．

本稿では，エージェントを学習させるために複数のエピソードを用いる．エージェントはエピソードの最後に至るまで連続して行動を実行し，一連の状態を移行していく (本稿のモデルでは，エピソードは 24 エポックで構成される)．エピソードの最後に，エージェントは行動を評価し，ポリシーを更新するための報酬を受け取る．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 状態の集合

最適なバッテリ残量レベル B_0 は文献 [8] のモデルを用いることで統計的に決定することができる．本稿では，この B_0 をエネルギー中立状態の規準とし，実際のバッテリレベルとの差分を求めるために利用する．エポック t_k でのバッテリ残量レベルを $B(t_k)$ とすると，エネルギー中立状態からの差分 $e_{dist}(t_k)$ は式 (7) で計算される．

$$e_{dist}(t_k) = B(t_k) - B_0 \quad (7)$$

e_{dist} は，エネルギー中立状態からの差分を表す状態 $S_{dist}(t_k)$ を決定するために用いられる．この情報を状態定義として用いることで，実際のバッテリ容量にほぼ依存せずに我々のアプローチを汎用的に利用することができるようになる．しかし，エージェントは，バッテリー残量が過剰に充電される，あるいは完全に放電されるというような危険領域にあるかどうかを考慮する必要がある．これはバッテリ残量レベルの状態 $S_{batt}(t_k)$ で反映される．状態 $S_{day}(t_k)$ は，エージェントが予期する天気の種類を示すものであり，この状態の情報を利用することで，エネルギー中立オペレーション達成に向け，エージェントは発電電力量の予測に依存して異なる戦略を採ることが可能になる．現在のエポック中に環境発電により得られた電力量の状態は $S_{eharvest}(t_k)$ で与えられる．最終的に，エージェントが取り得る状態は $S_{batt}(t_k)$，$S_{dist}(t_k)$，$S_{eharvest}(t_k)$，$S_{day}(t_k)$ の組み合わせとなる．すなわち，

$$(S_{batt}(t_k), S_{dist}(t_k), S_{eharvest}(t_k), S_{day}(t_k)) \in S$$

となる．

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，報酬は単に各エピソードの最後におけるエネルギー中立状態からの差分とする．各エピソード内での「行動の良さ」は，そのエピソード中の他の行動の影響も考慮しないと判定できない．そこで，選択された行動の良さの判定はエピソードの最後まで待ってから行われる．

各エピソードの最後で，その時のバッテリの残量レベルと初期のバッテリレベルの差が 0 であるようにエージェントがポリシーを学習することが理想である．そこで，差分が小さい場合には大きな報酬を，差分が大きい場合には小さな報酬を与えるようにする．

4.2 SARSA(λ)

あるエピソードの最後で得られた報酬は，Algorithm 2 [17] に基づいて状態–行動ペアの Q 値を更新するのに用いられる．適正度の履歴を用いることで，得られた報酬をそのエピソード中で報酬に貢献した全状態–行動ペアに伝搬させることができる．学習前に Q テーブルは大きな値である *qinit* で初期化される．各エピソードの最初に適正度の履歴をリセットし，初期の行動は 60%のデューティサイクルから開始される．その後，エージェントは自身の状態を得て，エピソードの最後まで ϵ-greedy ポリシーに基づいて環境と相互作用を行う．エピソードの最後で得られた報酬を利用し，Q テーブルをより良い行動推定ができるよう更新させつつ収束させる．

Algorithm 2 SARSA(λ) algorithm

```
Initialize Q(s,a) = qinit and e(s,a) = 0 for all s,a
for each episode do
    Initialize s,a
    for each step of the episode do
        Take action a, observe reward r and next state s'
        Choose next action a' from Q using ε-greedy policy
        δ ← r + γQ(s',a') − Q(s,a)
        e(s,a) ← e(s,a) + 1
        for all (s,a) do
            Q(s,a) ← Q(s,a) + αδe(s,a)
            e(s,a) = γλe(s,a)
        end fors ← s'; a ← a'
    end for
    until end of episode
end for
```

4.3 Q(λ)

Q(λ) 学習は，SARSA(λ) 学習と似通ったものであり，その違いは，Algorithm 2 のステップ 6 とステップ 7 である．ステップ 6 は省略され，ステップ 7 の式は以下のようになる．

$$\delta \leftarrow r + \gamma \max_a Q(s', a) - Q(s, a)$$

4.4 学習の設定

本稿では，エージェントの学習に過去のデータを用いる．ある年の日射量データを利用し，一日毎に SARSA(λ) を適用しつつ，その年のデータで N 回繰り返して学習を行う．学習には 3 つのフェーズを用いる．初期フェーズでは，バッテリレベルを中央付近の値に初期化する．2 番目および 3 番目のフェーズでは，それぞれ高い値および低い値に設定して学習を行う．これは，バッテリレベルが最適ではない状態から開始してもエネルギー中立オペレーションが達成できるようにさせるためである．本稿では，シミュレーションにより，学習時に経験していない環境下へとエージェントを設置した場合の振るまいを観測する．また，エージェントは，設置された環境へと適応化させるために，必要に応じて環境と相互作用しつつ学習を行い，自身の行動を調整する．

5. シミュレーションの方法

5.1 システムの仮定

本稿で用いるセンサノードは TMote Sky ノードの仕様を適切にスケールしたものを仮定する．TMote Sky は 3.6V で 2200mAh のリチウムイオンバッテリと，6W の太陽光パネル (220mm × 175mm) を備える．ノードの消費電力は，センシングや送信，受信を行っているかの状態に応じて，およそ 100mW から 20mW で変化する．本稿では，当該ノードの各値を約 5 倍にスケールアップした値を仮定する．

- **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 *100mWh* 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%, 40% ...100%*のように，*20%*刻みで設定可能とした．また，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 $B_{MAX} = 40000mWh$ で，理想的な特性を持つバッテリを仮定し，B_0 は B_{MAX} の *60%*に設定する．充電ロスや，漏れ電流など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り，特に考慮しない．

5.2 強化学習のパラメータ

5.2.1 行動の集合

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．本稿では $A = a(k) \in \{20\%, 40\%...100\%\}$ とする．

5.2.2 状態の定義

あるエポック t_k でのシステムの状態は次式で与えられる．

$$S_k = (S_{batt}(t_k), S_{dist}(t_k), S_{eharvest}(t_k), S_{day}(t_k))$$

エージェントの状態は，そのエージェントが実際に観測した値をもとに決められる．それらの値は一般的には連続値をとるため，各値がどの状態に対応するかを定める必要がある．本稿では以下のように各状態を定める．

$\boldsymbol{S_{batt}(t_k)} \in \{S_{b1}, S_{b2}, S_{b3}\}$: バッテリレベルが過充電，

表 1 $S_{batt}(t_k)$ の割り当て

$S_{batt}(t_k)$	Range
S_{b1}	$e_{batt}(t_k) < 20\%$ of B_{MAX}
S_{b2}	20% of $B_{MAX} \leq e_{batt}(t_k) < 80\%$ of B_{MAX}
S_{b3}	80% of $B_{MAX} \leq e_{batt}(t_k) < 100\%$ of B_{MAX}

表 2 $S_{eharvest}(t_k)$ の割り当て

$S_{eharvest}$	Range
S_{e1}	$e_{harvest}(t_k) = 0mWh$
S_{e2}	$0mWh < e_{harvest}(t_k) \leq 100mWh$
S_{e3}	$100mWh < e_{harvest}(t_k) \leq 500mWh$
S_{e4}	$500mWh < e_{harvest}(t_k) \leq 1000mWh$
S_{e5}	$1000mWh < e_{harvest}(t_k) \leq 1500mWh$
S_{e6}	$1500mWh < e_{harvest}(t_k) \leq 2000mWh$
S_{e7}	$e_{harvest}(t_k) > 2000mWh$

あるいは枯渇するなどの危険状態かどうかを与える．これは，$e_{batt}(t_k)$ の値に応じて表 1 により定められる．

$\boldsymbol{S_{dist}(t_k)} \in \{S_{d1}, S_{d2}, ..S_{d40}\}$: どの程度エネルギー中立オペレーションから離れているかを与える．本状態は $e_{dist}(t_k)$ の値をもとに以下のように決定される．

- $S_{dist}(t_k) \in \{S_{d22}, ..S_{d40}\}$: $e_{dist}(t_k) < 0$ の場合，すなわちバッテリレベルが B_0 未満の場合に対応する．
- $S_{dist}(t_k) \in \{S_{d21}\}$:完璧にエネルギー中立性が保たれた状態，すなわち $e_{dist}(t_k) = B_0 - e_{batt}(t_k) = 0$ に対応する．
- $S_{dist}(t_k) \in \{S_{d1}, S_{d2}, ..S_{d20}\}$: $e_{dist}(t_k) > 0$ の場合，すなわちバッテリレベルが B_0 より上である場合に対応する．

なお，状態は 1000mWh の間隔で設定される．

$\boldsymbol{S_{eharvest}(t_k)} \in \{S_{e1}, S_{e2}...S_{e7}\}$: エポック (t_k) において，どの程度のエネルギーが発電できたかの状態を与える．これは，$e_{harvest}(t_k)$ の値に応じて表 2 により定められる．

$\boldsymbol{S_{day}(t_k)} \in \{S_{f1}, S_{f2}...S_{f6}\}$: エージェントが予期するその日の天候の状態を表 3 に基づいて与える．実際には，天候情報は天気予報アプリケーションや天気予報情報が記載された Web サイトから取得される．本稿では，e_{day} の値により，日毎の天候状態を 6 つの状態に区分する．e_{day} は，その日に発電される総エネルギーであり，以下の式 (8) により求められる．

$$e_{day} = \sum_{k=1}^{24} e_{harvest}(t_k) \tag{8}$$

5.2.3 報酬関数

用いる報酬関数を図 2 に示す．図 2 は，e_{dist} と報酬の関係を示している．

5.3 学習パラメータ

評価では，学習パラメータとして $N = 10000$，$\alpha = 0.1$，$\gamma = 0.8$，$\lambda = 0.8$ の値を用いる．これらは，他の強化学習を利用した問題でもそうであるように，経験的に求めたものである．また，各エポックは 1 時間であり，1 エピソードは 24 エポックで構成される．エージェントは，東京地点の 2010 年の日射量データから表 4 に示す特定の日を用いて学習を行う．表 4 には，理論的に求まるその日の最良の平均デューティサイクルも示している．

表 3 $S_{day}(t_k)$ の割り当て

S_{day}	Weather	Range
S_{f1}	Very little sun	$e_{day}(t_k) < 2500mWh$
S_{f2}	Overcast	$2500mWh \leq e_{harvest}(t_k) < 5000mWh$
S_{f3}	Partly Cloudy	$5000mWh \leq e_{harvest}(t_k) < 8000mWh$
S_{f4}	Fair	$8000mWh \leq e_{harvest}(t_k) < 10000mWh$
S_{f5}	Sunny	$10000mWh \leq e_{harvest}(t_k) < 12000mWh$
S_{f6}	Very Sunny	$e_{harvest}(t_k) \geq 12000mWh$

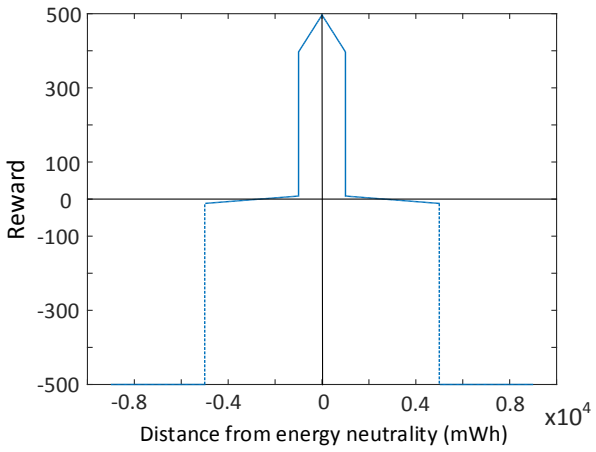

図 2 報酬関数

表 4 学習に用いた日にち

DAY	Total Energy Received (mWh)	Best Duty Cycle
265	13296.25	110.80%
80	11990.00	99.92%
101	10800.63	90.00%
37	9625.00	80.21%
69	8415.00	70.13%
343	7218.75	60.16%
329	6050.00	50.42%
53	4716.25	39.30%
277	3575.00	29.79%
61	2433.75	20.28%
102	1244.38	10.37%
303	515.625	4.30%

6. 評価結果

本章では, SARSA(λ) ポリシー, SARSA ポリシー, Q(λ) ポリシー，Q 学習ポリシーの 4 つの強化学習アルゴリズムで得られた各ポリシーの性能の比較評価を行う．評価では，東京地点の 2010 年，および 2011 年の日射量データを用いる．また，各ポリシーのエネルギー中立オペレーションの効率を評価するために，バッテリレベルは毎日 B_0 に

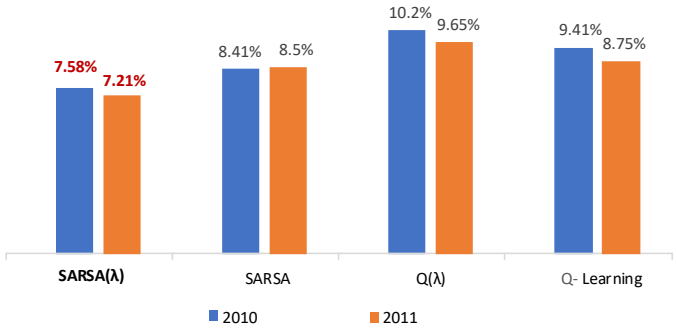

図 3 Net Energy Neutral Performance

初期化し，その日の終了時点までの性能を評価する．

本評価では，*Net Energy Neutral Performance* (NENP) と *Optimal Hit Percentage* の 2 つの指標を用いて比較評価を行う．ある日の NENP は，その日の終了時点でのエネルギー中立オペレーションからの差分である．NENP が低い値となるポリシーが良いポリシーとなる．NENP は B_{MAX} に対する比率で表すものとする．Optimal Hit Percentage は，ある年に，各ポリシーが他のポリシーと比較して最良の性能が得られた回数の割合である．

6.1 NENP

図 3 に，2010 年および 2011 年における平均のエネルギー中立オペレーションからの差分を示す．図より，SARSA(λ) ポリシーが最小の差分を達成し，続いて SARSA ポリシーが良い結果になっていることがわかる．SARSA アルゴリズムにより，適正度の履歴を用いる用いないに関わらず，Q 学習に比べて良い結果を得られている．この性能の違いは，式 (4) と式 (5) の両者のわずかな違いからきたものである．

SARSA 手法では，適正度の履歴を用いることで性能が向上する一方で，Q 学習手法の場合にはそうではないこともわかる．これは，高い α の値を用いたためと考えられる．高い α の値を用いると学習時にエラーの影響も大きく伝播されるためである．

Q 学習手法は，Q テーブルは既に最適であると仮定して，greedy ポリシーに基づいて行動を選択するが，実際には最適でない場合もある．Q 学習の場合，実際は ϵ-greedy 手法を用いているが，Q テーブルを greedy ポリシーに基づいているとして更新する．一方で，SARSA 手法では Q テーブルを更新することを考慮して更新が行われる．このわずかな違いが SARSA と Q 学習の性能の違いに表れていることは興味深い結果である．

全体として，Q(λ) ポリシーが比較対象の中で最も悪いポリシーとなっている．これは，学習率の設定によるものと考えており，高い学習率を用いているため，Q(λ) ポリシーでは最適な Q 値への収束が困難になっていると予想される．本稿では示していないが，学習率を 1/100 にすると，Q(λ) が優れた性能を示すことがわかっている．しかし，学習率を低くすることで，必要な繰り返しの回数が多くなってしまうという問題がある．

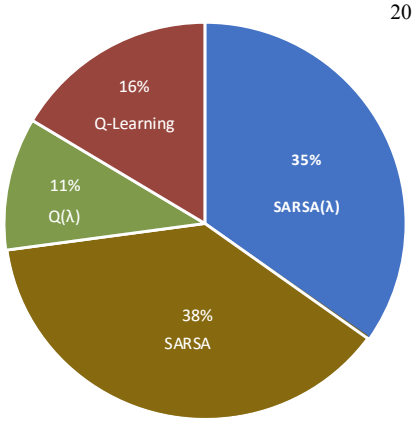

図 4 Optimal Hits (2010 年)

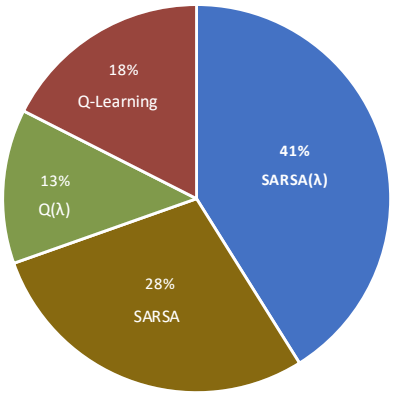

図 5 Optimal Hits (2011 年)

6.2 Optimal Hits

図 4 および 5 に，それぞれ 2010 年，2011 年の全ポリシーについての Optimal Hit Percentage の結果を示す．図より，SARSA ポリシーが最も大きな割合を占めていることがわかる．これは，概して SARSA，および SARSA(λ) が Q 学習手法を用いたポリシーに対して良いことを意味する．2011 年では，SARSA(λ) が 41%とかなりの部分の Optimal Hits Percentage を占めている．本評価結果は，前節の NENP の評価での結論の裏付けとなる．

また，2010 年には，SARSA ポリシーが SARSA(λ) ポリシーよりも高い Optimal Hit Percentage を達成していることは注目に値する．しかし，図 3 の結果からわかる通り，これは必ずしも良い性能が達成できることを意味していない．

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について，4 つの強化学習アルゴリズムの比較評価を行った．評価結果より，SARSA(λ) 手法をベースとした強化学習のアプローチが他のアプロー

チに比べて良いエネルギー中立オペレーションを達成できると結論付けることができる．適正度の履歴を用いることで SARSA アルゴリズムでは性能が向上したが，Q 学習アルゴリズムの場合は反対の結果となった．また， Q(λ) ポリシーは学習率の設定に影響を受けやすいことが観測された．そのため，Q(λ) ポリシーは長期間の，また時間のかかる学習には有効である一方，SARSA(λ) は少ない繰り返し回数で素早い学習が行えると言える．天気予報を利用の有無による性能比較や，他のパラメータを用いた評価などが今後の課題である．

謝辞 本研究の一部は JSPS 科研費 16K12405 の助成によるものである．

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] Gunduz, D., Stamatiou, K., Michelusi, N. and Zorzi, M.: Designing intelligent energy harvesting communication systems, *IEEE Communications Magazine*, Vol. 52, No. 1, pp. 210–216 (2014).
[4] Hsu, J. et al.: Adaptive duty cycling for energy harvesting systems, *Proc. of the 2006 ISLPED*, pp. 180–185 (2006).
[5] 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).
[6] 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).
[7] 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).
[8] Kansal, A. et al.: Power management in energy harvesting sensor networks, *ACM Tr. on Embedded Computing Systems*, Vol. 6, No. 4, p. 32 (2007).
[9] 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).
[10] Khan, J. A. et al.: Energy management in wireless sensor networks: a survey, *Computers & Electrical Engineering*, Vol. 41, pp. 159–176 (2015).
[11] 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).
[12] Michelusi, N. et al.: Energy management policies for harvesting-based wireless sensor devices with battery degradation, *IEEE Tr. on Communications*, Vol. 61, No. 12, pp. 4934–4947 (2013).
[13] Moser, C., Thiele, L., Brunelli, D. and Benini, L.: Adaptive power management in energy harvesting systems, *Proceedings of the conference on Design, automation and test in Europe*, EDA Consortium, pp. 773–778 (2007).
[14] Ortiz, A. et al.: Reinforcement learning for energy harvesting point-to-point communications, *2016 IEEE International Conference on Communications*, pp. 1–6 (2016).
[15] 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).
[16] Sudevalayam, S. and Kulkarni, P.: Energy harvesting sensor nodes: Survey and implications, *IEEE Communications Surveys & Tutorials*, Vol. 13, No. 3, pp. 443–461 (2011).
[17] Sutton, R. S. et al.: Reinforcement learning is direct adaptive optimal control, *IEEE Control Systems*, Vol. 12, No. 2, pp. 19–22 (1992).
[18] 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).
[19] Xiao, Y. et al.: Bayesian reinforcement learning for energy harvesting communication systems with uncertainty, *2015 IEEE International Conference on Communications*, pp. 5398–5403 (2015).
[20] シュレスタマリ，近藤，中村：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，情報処理学会研究報告 2017-ARC-225 (26) (2017).