知的ネットワークシステムへの強化学習の適用

- Q-learning による知的照明システムの構築 -

Application of Reinforcement Learning to Intelligent Network Systems
- Construction of Intelligent Lighting Systems by Q-Learning -

冨田 浩司
Koji TOMITA

Abstract: In this paper, we improve the control part of the intelligent networking systems. Our proposed intelligent networking system is the autonomous and distributed systems, and it is constructed with the intelligent artifacts that should have three components: sense, judge and act part. The whole network system has the same purpose and each intelligent artifact try to satisfy the purpose by itself. we applied reinforcement learning(RL) to our proposed systems. Q-learning is the typical algorithms of the reinforcement learning. The simulation of lighting systems by Q-learning is constructed for making clear the effect of the improved system.

1　はじめに

近年，多くの機器は「知的」，かつ「ネットワーク化」の流れにある．今後，全ての機器がこの方向に進むであろうと考えられ，現在，それらの機器を使った様々なシステムの実用的な構成が検討されている．我々は 1 つとして，知的な機器を知的人工物という枠組みでとらえ，その知的人工物をネットワーク化する「知的ネットワークシステム」を提案し，研究を進めている．本システムは自律分散型であるため，各機器の制御をどのようにするかが重要となる..

本発表では，我々が提案している知的ネットワークシステムの各機器の制御に強化学習を適用することで，従来の手法よりも効率よく目的を達成することを示す．具体的には，強化学習の代表的な手法である Q-learning を用いた知的照明システムのシミュレーションによって，その有効性を検証する．

2　強化学習

2.1　強化学習とは

強化学習 (1) は，移動などの行為を行うエージェントが，教師付き学習 (Supervised learning) のような直接の教師を持たずに，行為に対する環境からの報酬だけから，適切（アルゴリズムによっては最適）な行為の学習を行う典型的な自律的学習である．報酬をもたらす行動を優先すべく「強化」することから，強化学習と呼ばれる．強化学習の枠組を Fig.1 に示す．

学習主体「エージェント」と制御対象「環境」は以下のやりとりを行う．

(1) エージェントは時刻 t において環境の状態観測 S(t) に応じて 意志決定を行い，行動 A(t) を出力

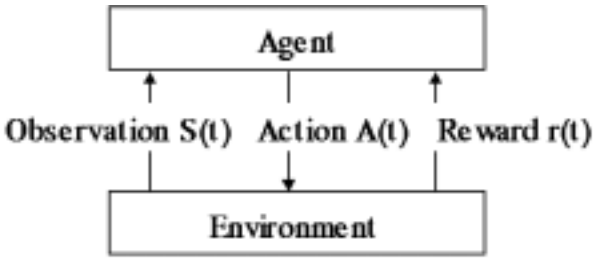

Fig. 1　強化学習の枠組み

(2) エージェントの行動により，環境は S(t+1) へ状態遷移し，その遷移に応じた報酬 r(t) をエージェントへ与える．

(3) 時刻 t を t+1 に進めてステップ 1 へ戻る．

強化学習では，設計者がゴール状態で報酬を与えるという形で，させたいタスクをエージェントに指示しておけば，ゴールへの到達方法はエージェントの試行錯誤学習によって自動的に獲得される．つまり，設計者が「何をすべきか」をエージェントに報酬という形で 指示しておけば「どのように実現するか」をエージェントが学習によって自動的に獲得する枠組となっている．これにより，制御プログラミングの自動化・省力化，ハンドコーディングよりも優れた解，あるいは自律性と想定外の環境変化への対応が容易に行える．

2.2　Q-learning

強化学習で最も代表的なアルゴリズムが Q-learning(2) である．Q-learning では，エージェントは状態認識器，行動選択器と学習器の 3 構成要素からなる．状態認識器は，状態と行動の対のテーブルすなわちルールベースで，各ルールは Q 値と呼ばれる重みを持っている．行動選択器には，Boltzmann 選択，ε -greedy 選択などが

あり，ボルツマン選択では exp(Q(s,a)/T) に比例した割合で行動を選択するのに対して，ε-greedy 選択はεの確率でランダム，それ以外は可能な行動の中で最大の Q 値を持つ行動を選択するなど様々なものが用いられる．学習器では次式に従って Q 値を更新する．

$$Q(s_t, a_t) := (1-\alpha)Q(s_t, a_t) + \alpha\,[r_t + \gamma\ \max Q(s_{t+1}, a)] \quad (1)$$

α は学習率，γ は割引率である．

あるスケジュールに従って学習率αを減少させ，多数の試行の後に Q 値が収束すると，各状態における最大の Q 値を持つルールの選択が最適な政策となることはすでに証明されている．

Q-learning のアルゴリズムを次に示す．

(1) エージェントは環境の状態 S(t) を観測する．

(2) エージェントは任意の行動選択方法（探査戦略）に従って，行動 A(t) を実行する．

(3) 環境から報酬 r(t) を受け取る．

(4) 状態遷移後の状態 S(t+1) を観測する．

(5) 上記の更新式により Q 値を更新する．

(6) 時間ステップ t を t+1 へ進めて手順 (1) へ戻る．

Q-learning の欠点は解析が保証しているのはあくまで最終結果であること，場合によっては非常に無駄な試行を伴い時間のかかること，学習の途中段階での Q 値は環境の構造や学習率などのパラメータに敏感であることが指摘されている．

3 知的ネットワークシステム [3]

3.1 知的ネットワークシステムの概要

我々が提案している知的ネットワークシステムは，近年における機器の「知的化」,「ネットワーク化」の流れに着目した具体的なシステムである．特徴は，ネットワークに接続する機器に知的人工物（詳細 次節）を用いること，自律分散型の基本技術を用いていること，各知的人工物それぞれが自律的に動作し，ネットワーク全体として与えられた大規模な目的を満たすことである．

まだ本システムは研究段階の領域を出ていないが，将来的には次のようなシステムを目指す．例えば，ある建物において「部屋を快適にしろ」という目的を与えると，接続されている知的照明，知的エアコンなどが例えば「部屋の温度を 28 度に維持し，人がいる所だけを明るくする」等を考え，自律的に部屋を快適にするシステム．また，交通システムにおいては，多くの交通機器（知的人工物）をネットワークに接続しておくことにより，例えば「交通渋滞を防げ」という目的を与えておくと，各交通機器はユーザからの命令を待たず，信号機故障や交通事故時による交通渋滞を解消するように自律的にネットワーク内で対処するシステム等である．

3.2 ネットワーク化される知的人工物

最近，特に家電製品や自動車によく見られる知的な機器を筆者の一人は「知的人工物」ととらえ．基本的な考察を行っている．[4]

知的人工物は必ず達成すべき目的を持つ．また，利用者を含む広義の環境条件の変化に対応して人工物自身のパラメータを自律的に変化させるために，その環境条件の変化をセンスするための各種のセンサ（認知）が必要である．次に，センサで得た情報を基に人工物の機能や性能を最適化する計画を立て（判断），それに沿って人工物のパラメータを変化させること（動作）ができなくてはならない．すなわち，全ての知的人工物は知的性質としてこの 3 つの要素を持ち，Fig.2 で表すことができると考えられる．

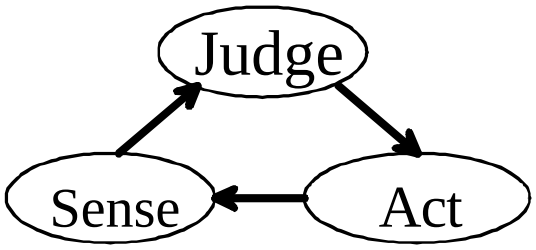

Fig. 2 知的人工物における知的構造

例えば，光感知照明機器は，外の明かりをセンスしあらかじめ組み込まれている明るさの判断基準から，光束を調節する知的人工物である．現在では知的とは言えない自動ドアも，人をセンスし，人の有無の判断基準から，ドアの開閉を制御するため，知的人工物の一つであると考えられる．

知的ネットワークシステムでは，ネットワークに接続する知的人工物が人工知能を持つ必要は決してない，後から加える必要もない．この考察からわかるように，知的人工物が必ず持つ「知的性質」を利用して動作させるため，既存の知的人工物をそのままネットワーク化した構成となる．

3.3 知的ネットワークシステムの動作手順

ネットワーク化する意味は大きくわけて次の 2 つである．まず，既存の各知的人工物をネットワークに与えられた目的を満たすように動作させるには，既存の目的を変更する必要がある．そのため，新しい目的を送るためにネットワーク化が重要となる．また，一つの知的人工物のセンサでは限界があるため，他の知的人工物の各種センサから得られる情報も必要である．そのため，情報の交換を行うためにネットワーク化が重要となる．

次に本システムの流れを示す．まず，各知的人工物は接続されるとネットワークに与えられた新しい目的を取り込む．この時点で，既存の判断基準は意味を失う．そのため，新しい目的に合った判断基準をそれぞれが生成する．各知的人工物のセンス部では各種環境をセンスすると同時に，他の知的人工物からのセンス情報も取り込み，取得した目的，各センス情報を基に次の制御を決定し，目的を満たすように各知的人工物が動作する．

3.4 知的人工物ネットワークシステムの有効性

本システムの有効性として，自律分散型を基本技術を用いているため，機器のネットワークへの参入・離脱が容易である，機器の故障によるシステム全体の停止を防ぐ，フレキシブルなシステムの拡張性が挙げられる．さらに，与えられた「目的」に対して，ネットワークに接続されている知的人工物だけで満たすように動作できるため，次のような有効性がある．

- 1 つの機器では不可能な作業を行うことができる．
- ある機器の故障時に起こる機能低下を他機器によって柔軟に対応し，補うことができる．
- 人間が考えつかないような動作で目的を満たすことができる．

4 知的照明システム

4.1 知的照明システムの概要

知的照明システムは知的ネットワークシステムの一つであり，本システムの基礎的な検討を行うのに使用している．システム構成は fig.3 に示すように，複数の知的照明機器（以下 知的照明）をネットワークに接続し，ネットワークに与えた「人がいる所を X [lx] の明るさにせよ.」という共通の目的を満たせるかどうかを検証する．

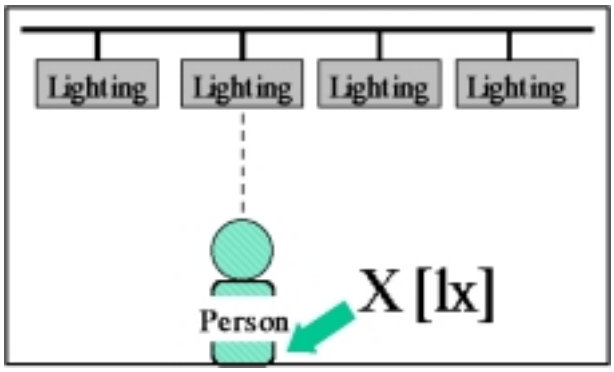

Fig. 3 知的照明システム

ここで用いられる知的照明は人感知センサと明るさ感知センサの両方が備わっているものとし，各知的照明は，各真下の人の有無と明るさ [lx] をセンスし，人の有無に合わせて調光することができるタイプを用いた．各知的照明の調光パターンは光度 0〜1,000[cd] とした．

各知的照明の従来の制御アルゴリズムを以下に示す．

(1) 各知的照明は一斉に，各々一度だけランダムに動作してみる（± 10[cd]）．

(2) 人の真上にいる知的照明は，(1) 後の環境（人がいる場所の照度）をセンスし，その情報をネットワーク全体に送る．

(3) 各知的照明は (2) の行動によって，目的への達成度が上がったかどうかを判断する．上がったならば，各知的照明はもう一段階上の動作を行う．下がったならば，再度 (1) の動作を行う．

(4) この手順の繰り返しにより，他の情報，自分の動作の有効性がわからなくても，知的照明全体で目的を満たすように動作することができる．

4.2 Q-learning を用いた知的照明システム

知的ネットワークシステムにおいて最も困難な問題は，各種センサ，情報はある程度あるが，それをどのよう使えば各知的人工物が目的に満たすような動作するのかである．つまり最適化手法の問題である．前節で従来の制御アルゴリズムを述べたが，判断基準を自動生成するわけではなく，予め与えている．そのため，柔軟性はなくこれ以上の機能は発揮できない．しかしながら，強化学習である Q-learning は目的（報酬がもらえる場所）さえ明確であれば，判断基準を自動生成してくれるため，本システムにおける最適化手法として，最も適していると考えれる．

そこで，知的照明システムに Q-learning を用いた．Q-learning を用いた知的照明システムのアルゴリズムを次に示す．

(1) 人の真上にいる知的照明は現在の環境（人がいる場所の照度）状態 S を観測し，他へ送る．

(2) 各知的照明は行動選択方法の一つである Boltzmann 選択に従って光束を強めるか弱めるかを決めて，行動 A を実行する．

(3) 各知的照明は報酬 r を受け取る．

(4) 人の真上にいる知的照明は次の環境（人がいる場所の照度）状態 S を観測し，他へ送る．

(5) 各知的照明はそれらの情報を基に (1) 式により Q 値を更新する．

(6) この手順を繰り返す．

4.3 シュミレーション

シミュレーションでは，1 つの知的照明では不可能な明るさを「目的」とした場合に，各知的照明が協力して「目的」を満たせるかを従来の手法を用いた知的照明システムと Q-learning を用いた知的照明システムで比較を行った．

Table 1　パラメータ設定

目的照度	150[lx]
誤差	5[lx]
状態 S の数	60 状態 (S0〜S59)
行動 A の数	2 状態 (A0，A1)
各 Q 値の初期状態	0.1(全て共通)
学習率	α =0.5
割引率	γ =0.9
行動選択方法	ボルツマン選択 (T=0.2)

各パラメータは次のように設定した．状態 S は 5[lx] 単位で 60 状態に分割し，S0(0〜5[lx])〜S59(195〜300[lx]) とし，各状態 S における行動 A は 2 状態，A0(+10[cd])，A1(-10[cd]) とした．学習率，割引率等は，予備実験による経験的な知見を参考に設定した．

Fig. 4　シュミレーション結果例

Q-learning を用いた場合のシュミレーション結果の一例を Fig.4 に示す．また，Fig.5 および Fig.6 にそれぞれの目的達成までの軌道を示す．

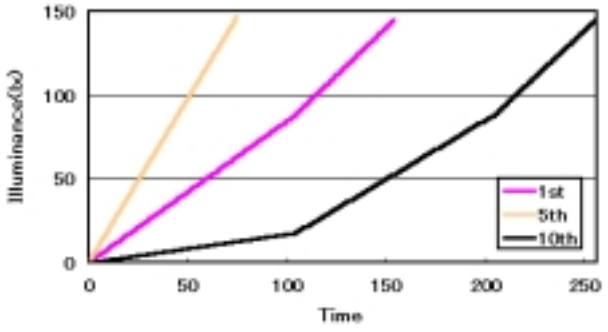

Fig. 5　従来の知的照明システムの軌道

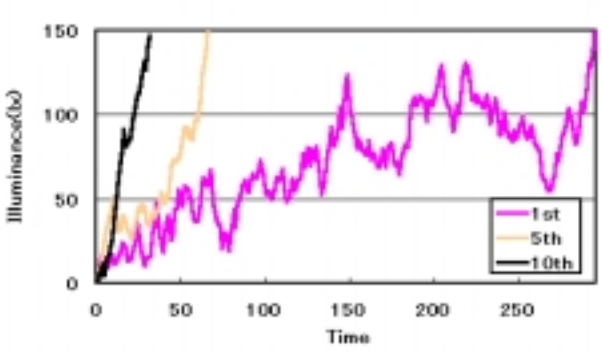

Fig. 6　Q-learning を用いた知的照明システムの軌道

横軸は目的を達成するまでにかかった時間であり，縦軸は合計照度である．各知的照明が消えている状態から目的が達成されるまでを 1 試行とし，表には 1 試行目，5 試行目，10 試行目のみを示した．

結果からわかるように，従来の知的照明システムは設計者が予め与えておいた制御で動作するため，目的を満たすまでの時間が試行回数に関わらず不安定である．一方，Q-lerning を用いた知的照明システムは試行回数を重ねるほど学習していき，10 試行目には目的を達成するのに 50 ステップかかっていないことがわかる．

5　結論と今後の課題

本発表では，知的ネットワークシステムの一つである知的照明システムにおいて，各機器の制御に代表的な強化学習である Q-learning を用いることで従来の手法よりも効率よく目的を達成することをシミュレーションによって検証できた．

また，従来の手法では判断基準を設計者が予め与えているのに対して，Q-learning では判断基準が自動的に生成されるため，異なった機器を接続した場合でも柔軟に適応できると考えられる．

参考文献

1) 畝見,『強化学習』，人工知能学会誌，pp.830-836(1994)

2) Watkins,C.J.C.H.and Dayan,P:Techical Note:Q-Learning,R.S.Sutton(ed.),Reinforcement Learning,pp.55-68,Kluwer Academic(1993)

3) 廣安，三木，冨田『知的人工物を用いた次世代ネットワークシステム〜知的照明システムによる基礎的検討〜』（日本機械学会:第 9 回設計工学・システム部門講演会，pp.518-521（1999））

4) M.Miki and T.Kawaoka『Design of Intelligent Artifacts:A Fundamental Aspects』（Proc.JSME International Symposium on Optimization and Innovative Design(OPID97)，1997-9）