知的照明システムにおける強化学習を用いた目標照度設定アルゴリズムの検討

中村 彰之

1　はじめに

本研究室では，ユーザの要求に応じて照明の制御を行う知的照明システムの研究を行っている [1)]．知的照明システムとは，任意の場所に任意の照度を提供するシステムである．現在の知的照明システムにおける目標照度の設定は，ユーザが直接的に照度などの物理量を操作している．しかし，知的照明システムの使用経験が浅いユーザにとって，物理量による要求は感覚的ではなく扱い難い．そのため，「とても明るく」や「少し暗く」といった感覚的な要求により，目標照度を設定することが望ましい．従って，ユーザからの感覚的な要求がどの程度の物理量に対応するか推測する必要がある．そこで，本研究では知的照明システムに機械学習を適用することにより，ユーザの感覚的な要求と物理量とを対応付ける．

機械学習には様々な手法が提案されているが，学習のために予備実験を行うものが多い．このような学習メカニズムを知的照明システムに適用させた場合，ユーザの負担が大きな問題になると考えられる．そのため，本研究では予備実験が不要な強化学習の導入を検討する．本稿では強化学習の概要と今後の研究の方向性について述べる．

2　強化学習

強化学習は人間の学習メカニズムを模倣している．人間の乳児が自然に物の掴み方を学習するように，システムが試行錯誤を繰り返すことで環境の状態に応じた適切な行動パターンを学習する．強化学習はロボットの行動制御の獲得やタスクスケジューリングなどに用いられている．

2.1　概要

強化学習では環境とエージェントという概念を導入している．環境は複数の状態を持つ対象問題であり，エージェントは対象問題で与えられた目的を果たすように行動規則の作成や行動決定を行う．自動車の運転を学習する人間に例えると，「安全に運転する」が目的に当たる．強化学習の流れを Fig. 1 に示す．

Fig. 1 において上部がエージェント，下部が環境を表している．また，矢印は環境もしくはエージェントの動作を示す．(1)〜(5) の各工程を繰り返すことにより学習を行う．Fig. 1 における各工程を以下に説明する．

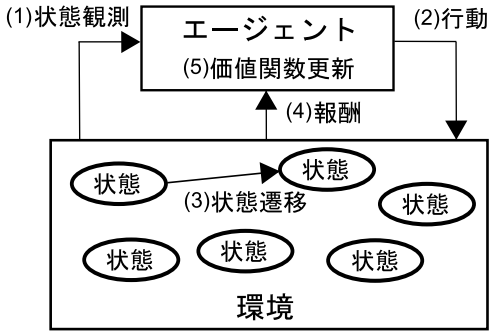

Fig.1　強化学習の流れ

(1)　状態観測
エージェントが現在の環境の状態を観測 (状態入力) する．自動車の運転の例では，「信号が赤である」や「人間が飛び出している」などが状態に相当する．

(2)　行動
エージェントは観測した状態や価値関数に応じて行動 (行動出力) を行う．ここで，価値関数とは，現在の状態や行動がどの程度良いのかを定義するもので，将来にどれほどの報酬が得られるかという期待値を表す．また，自動車の運転の例では，「ブレーキを踏む」や「ハンドルを切る」が行動に相当する．

(3)　状態遷移
エージェントの行動によって環境の状態が遷移する．

(4)　報酬
エージェントは行動の評価として，状態遷移に応じた報酬をスカラー値として環境から得る．行動によって状態が良くなればプラスの報酬を，悪くなればマイナスの報酬を与える．

(5)　価値関数更新
エージェントは得られた報酬を基に価値関数を更新する．

これらの工程を繰り返して価値関数を更新していくことで，より大きな報酬を得られる方策[*1]を獲得することが強化学習の目的である．そのため，価値関数の決定，更新方法が強化学習の性能を大きく左右する [2)]．

強化学習では，現在の状態と行動によって次の時刻の状態と報酬を予測可能にするため，マルコフ決定過程[*2]に従う．そのため，将来に得られる最終的な累積報酬 R_t は以下の式で与えられる．

[*1] ある状態におかれた際にどの行動を取るのかを決定するもの．
[*2] 将来における事象の起こる確率は過去のいかなる状態にも依存せず，現在の状態にのみ依存し決定する性質を持つ確率過程のモデル．

$$R_t = \sum_{k=0}^{T} \gamma^k r_{t+k+1}$$

ここで，r_t は離散時間 t における実報酬，T は最終時刻，γ は割引き率である．割引き率とは，遠い将来に得られる報酬ほど割引いて評価するものであり，$0 \leq \gamma \leq 1$ を満たす．

そして，この累積報酬 R_t を用いることにより，2 種類の価値関数が定義される．1 つは，状態 s にいることで将来に得られる報酬の期待値を示す状態価値関数 $V(s)$ である．もう 1 つは，状態 s 下で行動 a を選択することで将来に得られる報酬の期待値を示す行動価値関数 $Q(s,a)$ である．これらは，以下のように定式化できる．

$$V^{\pi}(s) = E_{\pi}\{R_t | s_t = s\} = E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\}$$

$$Q^{\pi}(s,a) = E_{\pi}\{R_t | s_t = s, a_t = a\}$$
$$= E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a\}$$

ここで，s_t，a_t は，それぞれ離散時間 t における状態，行動とし，π は方策，E は期待値を示す．方策 π とは，状態 $s \in S$ で行動 $a \in A(s)$ を取ることを意味し，$\pi(s,a)$ と表す．この 2 種類の価値関数は，強化学習の各手法により異なる [3) 4)]．

2.2 Actor-Critic

知的照明システムで扱う照度は連続値であるため，本研究では連続値の行動出力が可能な強化学習手法の 1 つである Actor-Critic の導入を検討する．

Actor-Critic は actor 部と critic 部により構成される．actor は方策を表現する構造を持ち，行動の選択に用いる．critic は価値関数を予測する構造を持ち，actor が選んだ行動を評価する．そのため，方策を表現する構造が価値関数から独立している．actor への評価には，TD(Temporal Difference) 誤差を用いる．TD 誤差とは，状態の評価が正しかったのかどうかを示すものである．Actor-Critic の概念図を Fig. 2 に示す．

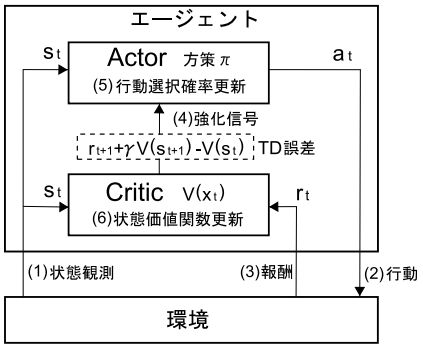

Fig.2　Actor-Critic の概念図 (参考文献 [2)] より参照)

(1)〜(6) の各工程を繰り返すことにより方策を決定する．以下に，Fig. 2 における各工程を説明する．

(1)　状態観測
エージェントが環境において状態 s_t を観測する．

(2)　行動
actor が方策 π に従って行動 a_t を実行する．

(3)　報酬
critic が報酬 r_t を受け取る．

(4)　強化信号
critic は次の状態 $s_t + 1$ を観測し，actor への強化信号として TD 誤差を計算する．TD 誤差 δ_t は価値関数 $V(s)$ を用いて以下のように表される．

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$$

(5)　行動選択確率更新
TD 誤差を用いて actor の行動選択確率を更新する．更新方法は後に解説する．

(6)　状態価値関数更新
TD 誤差を用いて critic の状態価値関数を更新する．

$$V(s_t) \leftarrow V(s_t) + \alpha(r_{t+1} + \gamma V(s_{t+1}) - V(s_t))$$

ここで，α は学習率と呼ばれ，$0 \leq \alpha \leq 1$ の係数である [2)]．

エージェントからの行動出力において，一般的な強化学習の手法は離散的な行動出力しか行えない．なぜなら，一般的な手法は価値関数により行動を選択するが，価値関数の構造が離散的だからである．それに対して，Actor-Critic は価値関数とは独立した部分で行動を選択する．そのため，行動を選択する actor が，乱数を用いて確率的な行動選択をすることで連続値の行動出力が可能となる．Fig. 3 を例に，actor での行動選択確率の更新を説明する．

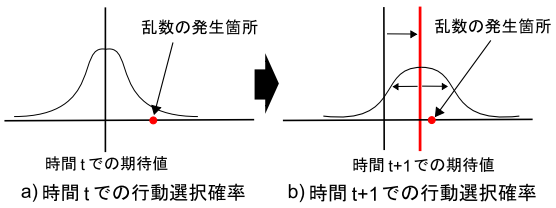

Fig.3　行動選択確率の更新 (参考文献 [5)] より参照)

Fig. 3 の図 a) と図 b) は，行動選択確率を更新した際の，時間 t(更新前) での行動選択確率と時間 $t+1$(更新後) での行動選択確率を表している．乱数発生箇所の値が示す行動 (行動 a とする) を行い，その行動 a の評価が良かった場合を想定する．更新前の行動 a は選択される確率が低い状況である．しかし，行動 a の評価は良かったため，より行動 a が選択され易くなるように行動選択確率を更新する．具体的には，図 b) のように，正規乱数の平均を乱数発生箇所に近付け，標準偏差の値を大きくする．このように行動選択確率を更新することで，actor は学習を行う [5)]．

3　数値実験

Actor-critic の学習の様子を確認するため倒立振子制御問題に Actor-Critic を適用し，数値実験を行った．本対象問題では，Fig. 4 に示すように台車上に倒立したポールのバランスを制御する問題であり，ポールのバランスを維持し続けることが目的である．

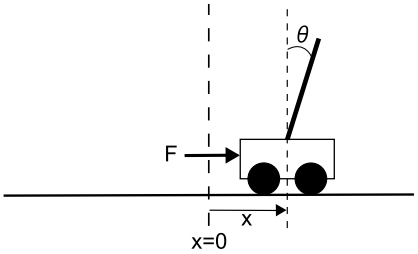

Fig.4　倒立振子制御問題 (参考文献 [6] より参照)

Fig. 4 において x は台車の中心地からの距離，θ はポールの傾いている角度，F は台車に加える力を示す．また，以下では $\dot{x}$ が台車の速度，$\dot{\theta}$ がポールの角速度となる．

台車の質量を $1.0kg$，ポールの質量を $0.1kg$，ポールの長さを $1m$，重力加速度を $9.8m/sec^2$ とした．このプログラムでは台車を押す力 F を actor の行動出力としたが，F は $10N$ もしくは-$10N$ を満たす離散的なものとした．また，環境からの状態入力を $(x, \dot{x}, \theta, \dot{\theta})$ とし，報酬はポールを倒した時にマイナスのスカラー値を与えた．

自作プログラムと参考文献 [7] のプログラムを性能比較した．100 回試行の平均値を Fig. 5 に，中央値を Fig. 6 に示す．

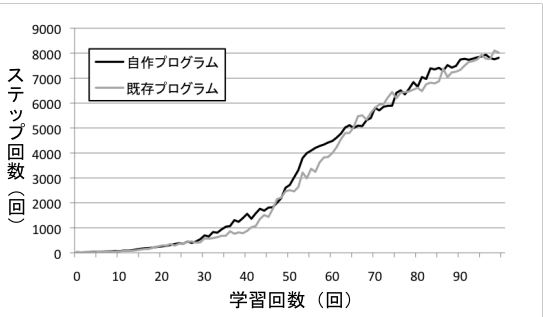

Fig.5　性能比較 (平均値)

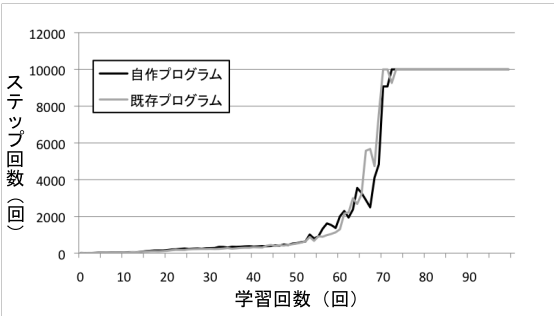

Fig.6　性能比較 (中央値)

Fig. 5 と Fig. 6 において横軸は学習回数，縦軸はステップ回数 (1 ステップ= 0.02 秒)，つまりバランスを保つことができた時間を示す．なお，この実験はステップ回数の上限を 10000 回とした．Fig. 5 と Fig. 6 の性能比較より，自作プログラムは既存プログラムと同等の性能を持っていることが分かる．

4　今後の研究

まず，Actor-Critic を知的照明システムに適用することにより，ユーザが望む照度変動量を知的照明システムに学習させる．この時，ユーザからの希望照度の要求を状態，知的照明システムへの照度設定を行動とする．また，報酬は以下の 3 つの与え方が考えられる．

- ユーザに報酬の入力を要求する．
- ユーザから照度変更の要求が無くなれば，ユーザの希望照度に設定できたものとして報酬を与える．
- ユーザから照度変更を要求されたなら，誤った照度に設定したものとしてマイナスの報酬を与える．

どの報酬の与え方が適しているかは，ユーザのシステムに対する使い易さや，学習の効率性の観点から検討を行う予定である．そしてパラメータの設定についても検討が必要である．学習の速度と精度が向上するような報酬の値や学習率などのパラメータの検討を行う．

次に，複数のユーザが利用する知的照明システムにおいて，ユーザ間で競合が起きる問題に着目する．この問題に対して，競合したユーザのそれぞれの目標照度を目的とし，複数の目的を同時に扱うことができる多目的最適化の概念を導入する予定である．そして強化学習を用いることで，ユーザ内での優先順位や照度の必要度などを考慮するアルゴリズムを構想している．例えば，競合したユーザ同士が上司と部下の関係であるなら，上司の目標照度を優先する．しかし，上司の照度の必要度が低い状況であるなら，部下の目標照度に幾分か近い照度を設定するというアルゴリズムの検討を予定している．

参考文献

1) 米澤基．知的照明システムのための自律分散最適化アルゴリズム．2005
2) 小林 (重) 研究室 - 強化学習
http://www.fe.dis.titech.ac.jp/research/rl/index.html
3) 強化学習の基礎
www.jnns.org/niss/2000/text/koike2.pdf
4) RWS MAIN -強化学習-
http://ryomiyo-web.hp.infoseek.co.jp/investigation/R_Learning/R_Learning.htm
5) Reinforcement Learning
http://mikilab.doshisha.ac.jp/dia/research/person/suyara/RL/
6) 木村元，小林重信．Actor に適正度の履歴を用いた Actor-Critic アルゴリズム：不完全な Value-Function のもとでの強化学習
7) Sutton & Barto Book: Reinforcement Learning
http://www.cs.ualberta.ca/~sutton/book/code/pole.c