状態集合価値関数を用いた時間推移対象向け強化学習手法の研究

State Action Set Based Reinforcement Learning for Time Dependent Events

若原 拓己[*1]
Takumi WAKAHARA

三上 貞芳[*2]
Sadayoshi MIKAMI

[*1] 公立はこだて未来大学大学院システム情報科学研究科 Future University Hakodate, Graduate School of System Information Science
[*2] 公立はこだて未来大学 Future University Hakodate

Abstract: Purpose of this research is developing method of reinforcement learning for objects that states changing related time shift. Nature environment especially plants growth is changing in time line (states changing related time shift), so in experiment of this research purpose is that new method applying to cultivation control in plant factory system.

1. はじめに

動的ネットワークにおけるパケットルーティングや，携帯電話の動的チャネル割り当てなど，強化学習手法の実用化が進められているが，リアルタイムで確率的にコントロールする制御であるという特性は，自然を相手にした制御に適していると考えられる．

しかし，近年の強化学習手法の研究は，状態の推移が時間推移の影響の無いロボットの制御などへ強化学習を適用させているものが主である．そこで本件旧では，従来あまり行われていない，植物育成などの状態が時間推移で行われるものについて，強化学習手法を適用させる事について考える．

2. 植物育成の制御

植物育成を工学的に制御するものとして，植物工場システムがある．植物工場システムでは，閉鎖的もしくは半閉鎖的な空間で，植物およびそれに付随する生物などを計画的，合理的に生産する．温度，湿度，二酸化炭素濃度，光量といった植物に必要な環境は一部自然環境を用いる事があるが，多くの場合これらの環境は人工的に作りだされ管理される．多くの植物工場では土を使わない水耕栽培が行われている．

制御対象として扱う事が出来るのは，環境の各要素の制御や養液供給量の制御などである．本研究では植物の生育にもっとも大きく影響を与えると考えられる，養液供給を制御対象として用いる．ここでいう養液とは，窒素，リン，カリウムを主成分とした液肥を含んだ水溶液の事である．

植物の成長度合いの指標としても様々なものが考えられる．植物の高さ，葉の大きさ，葉の茂り具合，果実を持つものは糖度などが考えられる．本研究の実験例では，計測のしやすさの点および育成対象を葉ダイコンとした点を考慮して，植物の高さを成長度とし，これを強化学習が用いる状態の指標として扱う．

3. 状態集合価値関数を用いた時間推移対象向け強化学習手法

植物育成と言った自然環境を対象として強化学習手法を適用させる場合，強化学習手法で用いる状態というのは，その推移が時間推移で行われるものとして本研究では扱っている．

本研究で提案する手法は，対象がある一定の基準に達した時点で，強化学習を行うための状態を推移させ，強化学習に使われる行動出力を，状態変化までの時間内で実際にあった行動出力の集合として扱う．価値関数の更新式を以下に示す．

$$Q(\{(s_i,a_i)\}) \leftarrow Q(\{(s_i,a_i)\}) + \alpha\left(r + \gamma\, Q(\{(s_i',a_i')\}) - Q(\{(s_i,a_i)\})\right)$$

状態 s_i から s_i' に推移する時間が t であり，行動出力が一定の間隔 Δt で行われたとすると s_i ，a_i は以下のように表される．

$$s_i = \{s_1, s_2, ..., s_n\}, a_i = \{a_1, a_2, ..., a_n\}, n = \frac{t}{\Delta t}$$

上記のように，行動出力を集合として扱う理由について，植物育成の場合，観測における強化学習で扱う状態と，制御上での状態のあり方が違うためである．特に報酬の与えられるタイミングが違ってくる事が大きな理由となる．

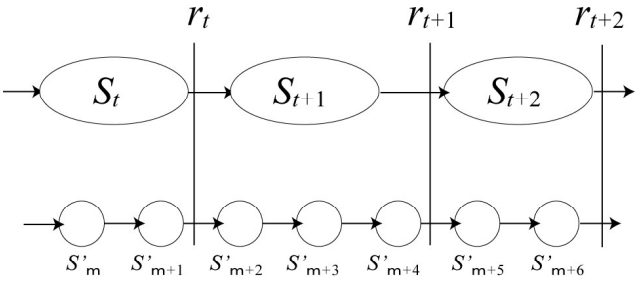

Fig.1 States about measure and control, Different timing of get rewards

Fig.1 の上段が観測上の学習系列で，下段が制御上の学習系列である．実際に報酬が与えられるのは，観測によるもので，制御上での学習系列で得られる報酬というのは，それと同じタイミングで得られる．従来の強化学習方式で問題となるのは，特に価値関数の更新を行う時である．例として観測上の状態 s_{t+1} と制御上の状態 $s'_{m+2} \sim s'_{m+4}$ について考える．観測上で得られる報酬はこの場合 r_{t+1} であり，従来の強化学習方式の場合，価値関数の更新が行われるのは s'_{m+4} のみで，s'_{m+2} および s'_{m+3} については報酬が与えられないので，学習が繰り返されたとしてもこれらの状態の価値関数は更新が行われることは無い．提案する手法，つまり状態を集合として扱った場合では，観測上の状態 s_{t+1} は制御上の状態 $s'_{m+2} \sim s'_{m+4}$ の集合として考えるので，s_{t+1} について価値関数の更新を行う，つまり $s'_{m+2} \sim s'_{m+4}$ について報酬を分配して与える事により，不足な

連絡先:若原 拓己，公立はこだて未来大学院 システム情報科学研究科， 〒041-8655 北海道函館市亀田中野町 1116 番地 2 三上研究室， g3107007@fun.ac.jp

く学習が行われると考えられる. この場合の報酬の分配方法としては, 観測上の学習系列での収益と, 制御上での学習系列の収益をほぼ同値にするために, 観測上の状態で集合として扱う, 制御上の状態群数で均等に分割した報酬を, 制御上の状態群それぞれに分配するのが妥当だと思われる.

従来手法と提案手法を比較すると従来手法では, 制御上の状態に関して, 価値関数の更新が行われない, つまり学習が行われない状態が発生してしまう. 対して提案手法では不足無く学習が行われるため, 従来手法に比べ効率的に学習が行われると考えることが出来る.

4. 小規模植物工場システムにおける実験

4.1 小規模植物工場システム

以下の図で示される小規模植物工場システムを作成し, それを用いて実験を行った.

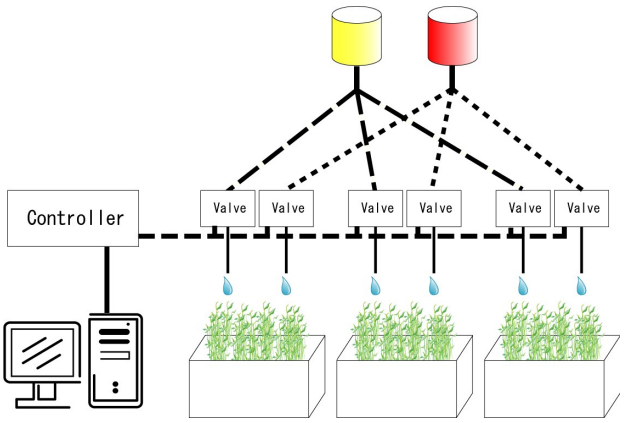

Fig.2 Small scale plant factory system

実験では, 一回の育成で葉ダイコンを同時に 3 株 6 日間育成させ, 各株の長さの平均を大きくすることを目標に, 2 種類の異なる成分比を持った養液(ハイポネックスハイグレード栄養素強化 064, 同 744)の供給比率の決定について学習を行う. 実験環境は, 温度 25℃, 湿度 50%, 蛍光管ライトを 12 時間で on/off 切り替えで行う. 養液の供給は 1 日 2 回, 12 時間間隔で行う. 育成を一回の実験につき 3 株行うことにより, 一回の実験で 3 つの結果について学習を行わせる.

4.2 従来手法による実験

提案する強化学習手法による実験の前に, 従来の強化学習手法による実験を行った. 実験の一回目は初期段階であり, 各状態の行動は価値を持っていない. そのため, 養液の供給比率がランダムなものを 2 株, 養液の供給比率が同一のものを 1 株とした. それ以降は強化学習を行い, プログラムにより養液の供給比率を決定させるようにした. 強化学習で扱う状態は養液供給を行う各タイミング, つまり 12 時間で 1 状態である. 行動出力に関しては 9 パターンとし, 養液供給比率の決定である. 報酬は収穫時, つまり 1 試行が終わった後に得る. 収穫した時点で各株の葉ダイコンの長さを計測し, その平均値を報酬とした. 学習の目標は 1 株あたりにおける葉ダイコンの高さの平均値を大きくすることである. 結果を Fig.3 に示す.

結果について, 一回目に比べ 2 回目, 3 回目は改善された結果を得ることが出来たが, それ以降は改善されている結果とは言いがたい. このような結果が得られたのは, 従来の強化学習手法では, 制御上での学習系列における各状態行動対の価値が適切に更新されないことが理由で, 適切な学習結果が得られないためと思われる.

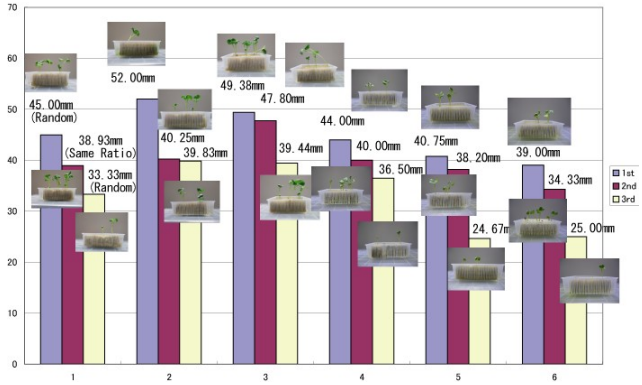

Fig3. Graph of result: ordinate is average of plants height, adscissa is episodes.

4.3 提案手法による実験

提案手法による実験について, 一回の実験で 12 回行う養液供給に関し, 前半 6 回, 後半 6 回に分けそれぞれのタイミングで観測を行い, その観測結果を報酬とする. 得られた報酬は観測までの状態数, この実験では状態数は前後半それぞれ 6 となるのでその数で均等に分割したものを報酬として与え, 各供給タイミングの行動出力の価値関数を更新する.

5. 結言

本研究では, 強化学習で扱う状態が時間推移に関係している対象向けの強化学習手法の提案を行った. 提案する手法について, 状態が時間推移と関係がある対象として, 植物育成に関し小規模植物工場を作成, それを用いた葉ダイコン育成の養液供給制御実験環境を作成した. その実験環境を用い, 従来の強化学習手法を用い, 2 種類の養液供給比率決定を行わせる実験を行った. 結果としては, 従来の強化学習に手法では回数を重ねるにつれ育成結果が改善されるという結果を得ることは出来なかった. 今後, 4.3 節であげた実験設定において, 提案した強化学習手法による育成実験を行っていく.

参考文献

[今 04] 今祐介: 自律移動ロボットの脱出行動戦略の適応学習, 日本機会学会全国大会講演論文集, 2004.

[高辻 93] 高辻正基: 植物工場の理論, SHITA TECHNOLOGY No.1, 1993.

[高辻 96] 高辻正基: 植物工場の基礎と実際, 裳華房, 1996.

[Sutton 00] Sutton, R.S., Bart, A.G. 著, 三上貞芳, 皆川雅章 共訳, 強化学習, 森北出版, 2000.

[Boyan 94] Boyan, J.A., Littman, M.L.: Packet Routing in Dynamically Changing Networks: A Reinforcement Learning method, Proc. NIPS94, 1994.

[Singh 96] Singh, S., Bertsekas, D.: Reinforcement Learning for Dynamic Channel Allocation in Cellular Tlephone, Proc. NIPS96, 1996.