

CELSS物質循環制御における複合制御手法の適用

Application of Composite Control Method for Material Circulation Control of CELSS

○中根 昌克, 広崎 朋史 (日本大学大学院)
塩澤 亮治 (NTT コムウェア株式会社 システム本部)
吉田 洋明, 石川 芳男 (日本大学)
宮嶋 宏行 (東京女学館大学)

Masakatsu NAKANE*, Tomofumi HIROSAKI*, Ryoji SHIOZAWA**
Hiroaki YOSHIDA*, Yoshio ISHIKAWA*
Hiroyuki MIYAJIMA***

* Nihon University, 7-24-1 Narashinodai, Funabashi, Chiba 274-8501 Japan

**NTT COMWARE System Integration Headquarters,

NTT Makuhari Bldg.1-6, Nakase, Mihama-ku, Chiba, Chiba, 261-0023, Japan

***Tokyo Jogakkan College, 1105 Tsuruma, Machida, Tokyo, 194-0004 Japan

*e-mail: m_nak@stone.aero.cst.nihon-u.ac.jp

Abstract

A Controlled Ecological Life Support System (CELSS) is essential for men to live for a long time in closed space such as a lunar base or a space station. A CELSS consist of much different characteristic equipment, so it is difficult to control the whole CELSS. To control the system, we regard those components as agents, and consider a CELSS as a Multi-Agent System. And we applied to Reinforcement Learning theory, to get its control law automatically. But, it occurred the problem that it exists the device which is not sufficient of learning. From the point that all agents need not to learn, we adopted "Composite Control" by Reinforcement Learning and other control theories. In this paper, we used a Rule-Based control as "the other control", and show the effectiveness of this Composite Control at system modification.

キーワード: セルス, 複合制御, マルチエージェントシステム, 強化学習

Key Words: CELSS, Composite Control, Multi-Agent System, Reinforcement Learning

1. はじめに

制御型閉鎖生態系生命維持システム (Controlled Ecological Life Support System: CELSS)とは、外部と物質のやり取りができない月面等の宇宙基地などの空間において、自律した物質再生循環系を確立するためのシステムである。その構成は大きく分けて、人間への食料供給や光合成により二酸化炭素を酸素へ再生するための「植物」、人間の生活によって出される尿・便・生活ゴミなどを処理し、また気体成分を調整するための物理・化学的な多くの「装置」、そして「人間」からなる。つまり、CELSSは非線形的な動きをする植物・人間・様々な動作を行う多くの装置から構成され、これらの要素間で複数種類の物質循環が行われる大規模で複雑なシステムとなる¹⁾。

我々はこのCELSSにおいて、安定した物質循環を維持することを目標に研究を進めてきた。宇宙空間における輸送の制限を考慮すると、宇宙における大規模システムは、小規模なものから出発して、しだいに増設・改良・修理等によって大規模化すること(システム変態)が考えられる。このようなシステムを従来の制御手法で構築しようとすると、各設備を独立して扱うことが困難なため、システム変態のたびに制御則を修正、あるいは再設計する必要が生じる。この問題を克服するために、我々は、CELSSを、各装置をエージェントとみなしたマルチエージェントシステム(MAS)として制御を行い、柔軟性・拡張性を獲得することを目指した²⁾。

システムをMASと見なし協調制御を行う際には、その協調方法を定める規約(協調プロトコル)を設計者が定めなければならない。しかし、本システムのように動的に拡張が行われるシステムでは、そのときの状況によりこのプロ

トコルは一般に変動してしまう。これに対処するためには、各エージェントが自律的に現在の状態を知覚し、適応していく機能が必要となる。そこで、我々は強化学習理論³⁾を適用し、協調プロトコルをエージェント自身が学習により自動獲得できるようにすることを目指した⁴⁾。

しかしながら系内には、連続運転を行うものから、バッチ処理を行うものまで、さまざま処理時間を有する装置が混在する。このような状況下で全ての装置に関して同様に学習を行わせると、バッチ処理を行う装置群のエージェントでは十分な学習回数を確保することができないという問題が発生した。このように、行動の意思決定時間がシステム要素により極端に変わってしまう場合、強化学習理論のみで制御するのは困難である^{5),6)}。

この問題に対処していく際、以下の点が重要となる。まずモジュールに関しては、その他の装置の状態に関わらず内部の環境状態を維持しなければならないため、モジュールで適用可能な制御方策は限定される。このため、システム変態が起こったとしても、モジュールに対しての制御方策は変わらない。次に長い処理時間をもつ装置に関しては、前述の通り、学習が十分に行われぬ。これは、このような装置に関係するエージェントに学習機構をもたせてもあまり効果がないことを意味している。この場合、始めから簡単なルールで制御を行っていたほうが良い結果を与える。これらのことから、システム変態に対処するために学習機能を用いるとしても、全てのエージェントに学習機能を付加する必要は無いといえる。

そこで我々は、システム変態に対応させることが可能な強化学習理論に加えて、装置の動作条件を考慮した制御方法を複合的に用いることを考えた。この具体的な方法とし

て、装置に関するエージェントに対してルールベース制御や予測制御を用いてシミュレーションを行ってきた。^{7), 8)}そこで、本稿ではその続報として、実際にタンク増設というシステム変態を起こし、その際にシステムとして破綻を起こしていないかを確認する。また、強化学習理論のみで同じシステム変態を起こした場合の結果と比較し、その違いに関して考察を行う。

2. 制御モデル^{4), 7), 8)}

2.1 CELSS モデル

CELSS のモデル化に際して、(財)環境科学技術研究所の閉鎖生態系実験施設 (CEEF: Closed Ecology Experiment Facilities) を参考にした。但し、研究初期段階であることを踏まえ、本稿では酸素・二酸化炭素の循環にのみ着目し、水・窒素・アンモニア・硝酸等の循環に関してはモデル化を行っていない。また、人間の活動モデル、および植物の成長モデルは参考文献¹⁾に従った。Fig.1 に、O₂・CO₂そして炭素 C の循環モデル図を示す。なお、図中における WOP・ORP は以下で説明する湿式酸化装置、酸素再生装置を表している。また、点線で表されているリンクはシミュレーション中に動的に拡張されるリンクである。

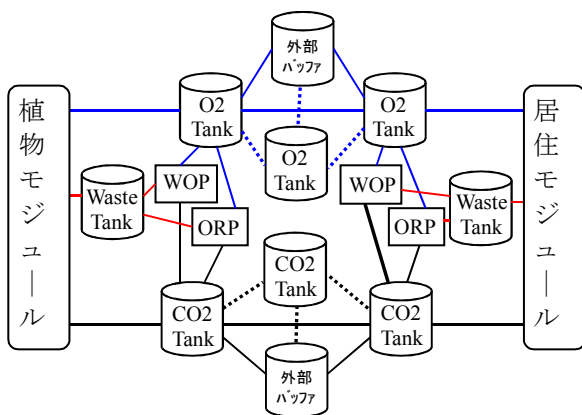


Fig.1 Circulation Model of O₂, CO₂, and C

以下に、本モデルにおいてバッチ処理を行う装置について概説する。

2.1.1 湿式酸化装置(Wet Oxidation Processor: WOP)

湿式酸化装置は、植物の不可食部あるいは人間の糞尿・生活屑などの廃棄物(Waste)を酸化分解することにより、有機物を二酸化炭素やアンモニアなどに変換する設備である。本稿では酸素と二酸化炭素の循環のみをモデル化しているため、この装置は酸素と廃棄物を吸収し、二酸化炭素を排出する。湿式酸化装置の処理は1バッチ480分かかる。また、1回のバッチ処理により反応する物質量は一定である。

2.1.2 酸素再生装置(O₂ Recovery Processor: ORP)

酸素再生装置は二酸化炭素を取り込み、還元・熱処理および電気分解を行うことにより酸素を生成する装置である。この際、副生成物として炭素が発生する。本モデルにおいては、二酸化炭素を吸収し、酸素と廃棄物を放

出する。酸素再生装置の処理は1バッチ60分である。また本モデルでは酸素再生装置を緊急時用と位置づけ、植物の光合成による酸素再生では間に合わなくなったときのみ作動させる。

2.2 危険レベル

危険レベルとは、各タンク内の物質量を幾つかの領域に分割し、離散的に表現するための指標である。この指標を基に3章で説明する制御行動を決定することになる。本研究では危険レベルを5段階で表現し、正の値を取ると目標値よりも多すぎることを表し、負の値を取ると目標値よりも少なすぎることを表す。すなわち、その絶対値が大きくなるほどタンクは危険な状態となる。Table 1 に危険レベルの定義を示す。

Table 1 Definition of Risk Level

タンク内の物質量		危険レベル
以上	以下	
最大許容量	最大量	2
(目標値+最大許容量)/2	最大許容量	1
(目標値+最小許容量)/2	(目標値+最大許容量)/2	0
最小許容量	(目標値+最小許容量)/2	-1
最小量	最小許容量	-2

ここで、最大量・最小量は各タンクの物理的な最大量・最小量を表す。また最大許容量・最小許容量は、その範囲内ならば通常運用を行い、この範囲を超えてしまった場合、非常事態になっていると判断する値である。

3. 複合制御手法⁸⁾

3.1 基礎概念

本研究では計算間隔を4分としてシミュレーションを行っている。このとき、連続運転ができるタンク(内蔵物質の輸送)では1日につき360回の学習ができる。それに比べ、バッチ処理を行う酸素再生装置は1日に最大24回、湿式酸化装置は最大2回のみしか学習の機会がない。これにより、学習の進行速度に大きな違いが生じ、バッチ処理装置群の学習不足が問題となった。そこで、この問題に対処するために本複合制御手法を考案した。

本複合制御手法は強化学習理論を用いた適応エージェントとルールベース制御を行う固定方策エージェントからなる。学習を行ってもその機会が極端に少なく、したがって効果が期待できないエージェントに対しては学習を行わず、人間の知見から得られるルールにより方策を固定する。一方、連続行動を行い、学習回数を十分に確保することができるエージェントに対しては強化学習理論を適用し、固定方策を採用するエージェントを考慮にいられた制御則を学習により獲得することを狙う。

以下、本稿で用いた強化学習理論とルールベース手法の詳細について説明する。

3.2 強化学習理論³⁾

強化学習理論とは、環境との相互作用を通じて、環境に適応する制御則を獲得する学習手法のことである。強化学習理論では、いわゆる教師信号とよばれる学習用信号を用意する必要がない。その代わりに、行動の結果から

得られる評価（報酬）を定義し、それを手掛かりとして、どの行動が最も高い報酬を得ることが出来るのかを試行錯誤を繰り返すことにより学習していく。Fig.2 に強化学習理論の概念図を示す。本研究では数ある強化学習理論のアルゴリズムの中から Q-Learning と呼ばれるアルゴリズムを採用した。

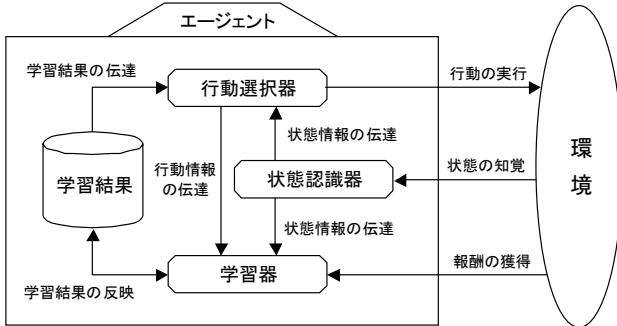


Fig.2 Image of Reinforcement Learning

3.2.1 Q-Learning アルゴリズム Q-Learning は強化学習理論のなかで最も一般的なアルゴリズムである。このアルゴリズムでは、 Q 値とよばれる、ある状態における、行動ごとの評価値を学習する。このことにより、その状態で行うべき最適行動を獲得することが可能となる。状態 $x \in X; X$ は現在の状態において知覚可能な状態の集合) において、行動 $a \in A; A$ は現在の状態において選択可能な行動の集合) をとった場合の Q 値は、 $Q(x, a)$ と記述される。エージェントが時刻 t の状態 x_t において行動 a_t を選択した結果、状態が x_{t+1} となり、報酬 r_t が得られたとすると、この Q 値の更新は以下のように行われる。

$$Q_{t+1}(x_t, a_t) = (1 - \alpha)Q_t(x_t, a_t) + \alpha(r_t + \gamma \max_{k \in A} Q_t(x_{t+1}, a_k)) \quad (1)$$

ここで、 α は学習率と呼ばれるパラメータであり、 $0 < \alpha \leq 1$ の値を設定する。また γ は割引率と呼ばれるパラメータであり、 $0 \leq \gamma \leq 1$ とする。

(1)式の右辺第1項は過去の学習による経験値を表している。また第2項は今回の学習結果（報酬）を表し、第3項は、未来の最適行動を採った場合の推定価値を表している。このように、次の状態において採用できる行動の価値を考慮にいたした学習をすることが可能である。

3.2.2 行動選択ルール エージェントは、学習経験が浅い段階ではランダムに行動を選択しつつ学習を行い、経験を積み十分な学習を行った後は学習結果を基に行動する必要がある。それらを実現するために、以下の方法を採用した。

- $\tau = \exp(-5 \times n_i / N)$ を計算する ($0 < \tau < 1$)。ここで、 n_i は現在の学習回数、 N は目標学習経験回数である。 n_i は、各状態について1つずつ定義する。
- $0 \sim 1$ の乱数を発生させる。その値が τ 以下の場合には、更に乱数を発生させてランダムに行動を選択する。 τ 以上の場合には、最大の Q 値をとる行動を選択する。

上記において、学習経験回数は状態 x_i 毎に定義しているため、定常状態から外乱等により非定常状態になった場合には、未経験の状態となるため、学習が促進される。従って、多種の外乱を経験させればさせるほど様々な状

態への対処が可能となる。

3.3 ルールベース手法

本稿のモデルにおいては、湿式酸化装置・酸素再生装置のエージェントがルールベース制御を行う。本モデルで採用されている湿式酸化装置・酸素再生装置は共に、装置の特性として1バッチで生成される物質量が固定されており、出力の制御は不可能である。しかしながら、これらの装置はタンクに蓄えられている物質の量によって行動が制御されるべきである。そこで、各装置が接続されているタンクの危険レベルを組み合わせ、組み合わせごとに行動確率を設定し、 $0.0 \sim 1.0$ の乱数を発生させその値が行動確率以下の場合には処理動作を開始し、そうでない場合はその時間での処理開始を見送る、という方法を用いた。以下で2つの装置に設定した制御ルールに関して説明する。

3.3.1 湿式酸化装置 酸素 O_2 と廃棄物から二酸化炭素 CO_2 を生成する湿式酸化装置に関しては以下のような制御ルールを設定した。

- 廃棄物タンクの危険レベルが -2 でない限り、処理は O_2 タンクと CO_2 タンクの危険レベルの組み合わせから算出する行動確率により行動を開始する。行動確率 P_w は知見より、以下の式(2)によって定めた。

$$P_w = 1.02401 - 0.2352 \exp(-1.141 \times Risk_{O_2}) - 0.2(Risk_{CO_2} + 2) \quad (2)$$

なお、 $Risk_{O_2}$ は O_2 タンクの危険レベル、 $Risk_{CO_2}$ は CO_2 タンクの危険レベルを表す。

- 例外として、 CO_2 タンクが空になりそうな(危険レベルが -2 である)場合、廃棄物タンクの危険レベルに関わらず、式(2)で行動確率を規定する。ただし、この場合、処理に必要な廃棄物量はタンク内に存在するものとする。

3.3.2 酸素再生装置 二酸化炭素 CO_2 から酸素 O_2 と廃棄物を生成する酸素再生装置に関しては以下のような条件を考慮する必要がある。

- この装置は廃棄物を排出するが、その量は無視できるほど微少である。そのため行動確率 P_o は O_2 タンクと CO_2 タンクの危険レベルの組み合わせで定めた。
- またこの装置は本モデルにおいては緊急時のみ作動させるという位置づけになっている。そこで、 O_2 タンクが空になりそうな(危険レベルが -2)場合と CO_2 タンクがあふれそうな(危険レベルが 2)場合に限り処理を開始するようにした。

これらの条件から Table 2 のような行動確率を設定した。

Table 2 Action ratio for ORP

$Risk_{O_2}$	$Risk_{CO_2}$	行動確率
-2	-2	0.000
-2	-1	0.288
-2	0	0.789
-2	1	0.949
-2	2	1.000
-1	2	0.800
0	2	0.400
1	2	0.000
2	2	0.000

これ以外の組み合わせに関しては行動確率=0.000

4. シミュレーション条件

複合制御手法の有効性を確認するため、CELSS モデルに複合制御手法を用いた物質循環のシミュレーションを行い、強化学習のみを用いたものと比較をおこなった。その際のシミュレーション条件を以下に示す。

- (1)植物として稲、ダイズ、レタス、トマト、さつまいも、ゴマを栽培する。栽培量は人間1人が生存するのに必要な量とする。植物の基礎データを Table 3 に示す。シミュレーション期間は800日間とし、人間は米の収穫後の120日目から滞在開始とした。また、シミュレーションにおける計算間隔は4分とした。

Table 3 Master data of plants

名前	栽培 日数	収穫 指数	収穫 量	栽培 面積	区画
	day	-	g/m ² / day	m ²	
稲	120	0.415	8	51	5
ダイズ	100	0.394	14	10.93	4
レタス	40	0.621	59.79	5.12	10
トマト	90	0.661	17	6	10
さつまいも	120	0.831	19	5.37	4
ごま	90	0.268	0.89	114.61	6

- (2)植物は、植物ごとに幾つかの区画に分離し、区画ごとに栽培・収穫時期をずらすことにより、食料供給量を安定化させる。
- (3)物質循環が安定するまでの期間はタンク容量をオーバーする場合があるため、外部バッファを用意した。すなわち、外部バッファを使用している期間は開放系での運用であり、外部バッファとの出し入れが無くなった時点から閉鎖系での運用となる。
- (4)植物モジュールの照明装置は、1日の中で12時間の明暗周期を持たせる。植物の成長度合は変動する。
- (5)強化学習のパラメータとしては、学習率 $\alpha=0.1$, 割引率 $\gamma=0.99$, 行動選択に用いる目標学習経験回数 $N=500$ とした。
- (6)各タンクの最大量・最大許容量・最小許容量・最小量は、Table 4 に示すとおりである。

Table 4 Master data of Tank

名前	最大量 [g]	最小量 [g]	最大 許容量 [%]	最小 許容量 [%]
PM 用 O ₂ Tank	4904.0	0.0	80.00	20.00
HM 用 O ₂ Tank	1854.0	0.0	80.00	20.00
PM 用 CO ₂ Tank	6125.0	0.0	80.00	20.00
HM 用 CO ₂ Tank	4446.0	0.0	80.00	20.00
PM 用 WasteTank	38611.0	0.0	50.00	2.569
HM 用 WasteTank	5000.0	0.0	50.00	16.00

PM: 植物モジュール, HM: 居住モジュール

- (7)システム変態の具体例として、シミュレーション開始から500日目にO₂タンク・CO₂タンクを増設し、Fig.1にて点線で表されているリンクを発生させた。

5. シミュレーション結果および考察

800日間の物質循環シミュレーションの結果を次ページ Fig.3~Fig.6 に示す。ここで、Fig.3 と Fig.4 は複合制御手法を用いた場合の結果である。また、対比するために、従来どおり強化学習理論のみを用いて、同じ条件にてシミュレーションした結果を Fig.5 と Fig.6 に示す。

まず、Fig.3 と Fig.4 から、提案している複合制御手法がタンク増設というシステム変態に対応できるかどうかを評価する。まず、酸素タンク群の時間変動を表した Fig.3 については、システム変体の前後で非常に安定した制御が実現されている。600日目において居住用タンクが一時的に危険レベル-2の領域に入ってしまった。これは、安定期においても強化学習理論が極小の確率でランダム行動により学習を行うために発生したと思われる。

次に、二酸化炭素タンク群の時間変動を表した Fig.4 について考察する。まず、シミュレーション開始初期に居住モジュール用二酸化炭素タンクが危険レベル-2の危険な状況になっている。これは、居住モジュール用二酸化炭素タンクが、人間が吐き出した二酸化炭素を植物モジュール用のタンクに送り込むことが可能となるときまで一時的に溜め込む役割であり、条件がそろえばすぐに物質の輸送を開始する。そのため、初期段階では低い値を取りやすいのが原因である。また、居住モジュール用タンクの容量が植物モジュール用に比べ少ないため、前記の理由が増幅されたと考えられる。しかしながら、その後はだんだんと欠乏状態から脱している。これは強化学習理論の再学習効果が上手く機能したためと思われる。500日目のシステム改変の後には、ほぼ安定な制御が行われている。唯一、600日目付近で一時的に物質量が低下しているが、これは、同時期に酸素タンクの危険レベルが-2、つまり欠乏に近い状態になっているために、酸素再生装置が作動したためである。このように、一部が学習を行わない複合制御手法でも十分にシステム改変に対応できることが示された。

次に、強化学習理論のみを用いておこなったシミュレーションの結果である Fig.5 と Fig.6 との比較を行う。Fig.5 と Fig.6 では、シミュレーション前期では安定した物質循環が行われている。しかしながら、タンクを増設が起こった後、内容量の振動が激しくなっている。これは、タンク拡張というシステム変態に対して、一時的にそれまでエージェント達が学習してきた制御方針と増設タンクが行う制御が競合しているためであると考えられる。特に、意思決定時間が長い湿式酸化装置や酸素再生装置に関連するエージェントに関しては、なかなか学習が行われず、その間は制御方針が安定しない。そのため、増設タンクに関連する

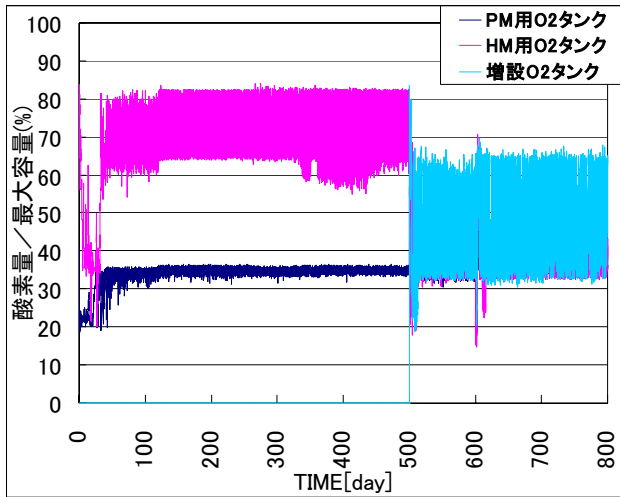


Fig. 3 Change in amount of material in O2 Tank (Composite Control)

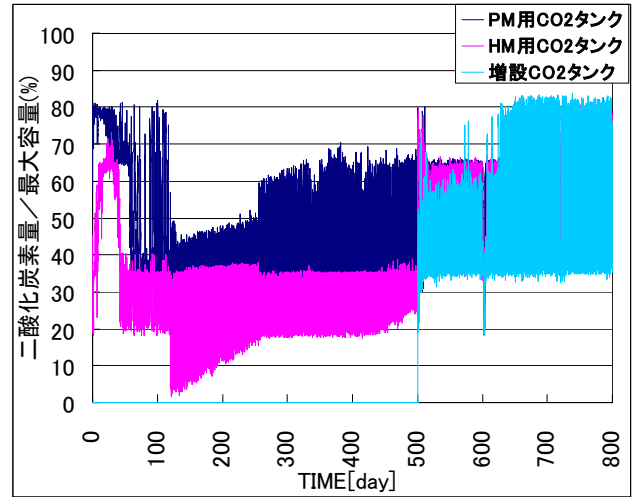


Fig. 4 Change in amount of material in CO2 Tank (Composite Control)

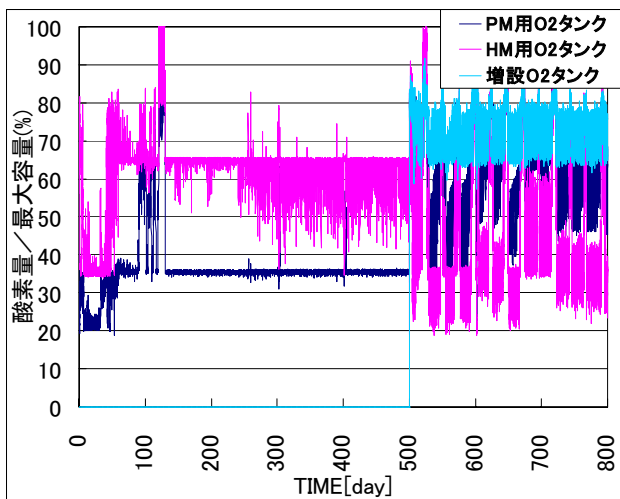


Fig. 5 Change in amount of material in O2 Tank (Reinforcement Learning Theory)

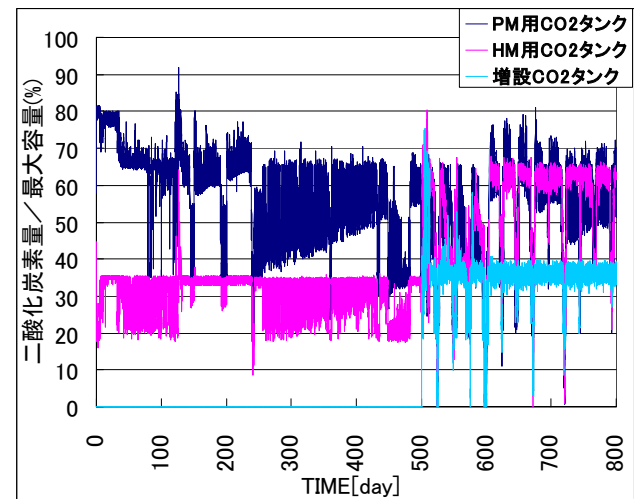


Fig. 6 Change in amount of material in CO2 Tank (Reinforcement Learning Theory)

エージェントが安定しない制御方針に悩まされ、その結果として制御が競合したと考えられる。また、増設したタンクは湿式酸化装置など意思決定時間が長い装置とは直接にはつながっていない。そのため、増設したタンクに関連するエージェントは時間遅れを伴って湿式酸化装置などが行う制御の影響を受けることになる。この点が、システム変態前後で大きく異なる点であり、前半では120日前後で安定しているものがシステム変態後はなかなか安定しない理由であると考えられる。

それに対して、複合制御手法を用いた結果である Fig.3 と Fig. 4 はシミュレーション全期において安定して制御されている。このように複合制御手法にてスムーズなシステム変態が行われた理由は以下のように考えられる。すなわち、意思決定時間の長い装置に関係するエージェントの制御方針をルールで固定化することにより、増設したタンクに関連するエー

ジェント群がはじめからその固定方針を考慮に入れた学習が行われたと考えられる。すなわち、前述した極小の確率で行われたランダム行動の影響で作動した酸素再生装置の影響で制御に微小な影響が生じただけの良い結果が得られたものと考えられる。

6. おわりに

今まで我々が提案してきた、強化学習理論のみで制御方針を獲得するという手法にあった問題点を、全てのエージェントが学習する必要がないという観点のもと、他の制御手法と複合的に用いることで解決しようと考えた。この手法に関しては、参考文献⁸⁾にて、理想的な状態での特性が考察されている。そこで本稿では、数値シミュレーションにより、タンク拡張というシステム変態が起きた場合でもシステムが破綻しないことを確認し、さらに、従来手法との比較の結果、従来手法よりも良い結果を得ることができた。

今後の展望として、以下のような点が考えられる。

- (1) 今回、強化学習理論と同時に用いる制御手法とし

て単純なルールに基づくルールベース制御を用いた。しかしながら、参考文献⁵⁾にあるように極端な学習回数⁵⁾の差に対処できる手法は他にも存在する。どのような制御手法が強化学習理論と相性が良いかを評価していかなければならない。

- (2) 今回は酸素・二酸化炭素・炭素の循環のみを考慮した簡単なモデルについて物質循環シミュレーションを行っている。しかし、モデルが代わればルールも再考する必要が出てくる。すなわち、モデルを厳密化した際、どのようにルールが変更されなければならないかを考察する必要があると考える。
- (3) 今回はシステム改変として、タンクの増設を取り扱ったが、それ以外にタンクの取り外しや物理化学処理装置の増設・取り外しなどが考えられる。これらはタンク増設時と系統の異なった問題となるので、まず問題点を明らかにしていく必要がある。

参考文献

- 1) 宮嶋宏行・石川芳男・芦田章・新田慶治：「シミュレーションモデルの開発と CEEF 統合試験への適用」. CELSS JOURNAL, Vol. 13, No. 1, 2000, pp. 1-11
- 2) 広崎朋史・安居院健治・石川芳男・宮嶋宏行：「CELSS 運用管理へのエージェント技術の適用に関する研究」, CELSS 学会学術講演会講演集, 2000,
- 3) 木村元・宮崎和光・小林重信：「強化学習システムの設計指針」. 計測と制御, Vol. 38, No. 10, 1999, pp. 618-623
- 4) 広崎朋史・塩澤亮治・中根昌克・吉田洋明・石川芳男・宮嶋宏行：「拡張性を有する CELSS 物質循環制御系構築手法」. 第 47 回宇宙科学技術連合講演会論文集, 2003, pp.516-521
- 5) 山田誠二 著・日本認知科学会 編集：「適応エージェント 認知科学モノグラフ 8」, 共立出版, 1997
- 6) 高玉圭樹：「マルチエージェント学習 ー相互作用の謎の迫るー」, コロナ社, 2003
- 7) 塩澤亮治・中根昌克・広崎朋史・吉田洋明・石川芳男・宮嶋宏行：「装置特性を考慮した協調制御手法及び強化学習による CELSS 物質循環制御」, 第 46 回自動制御連合講演会講演論文集, 2003, pp.370-373
- 8) 中根昌克・塩澤亮治・広崎朋史・吉田洋明・石川芳男・宮嶋宏行：「ルールベースと強化学習の複合制御を用いた CELSS 物質循環制御」, 第 16 回自律分散システム・シンポジウム資料, 2004, pp.79-84