知 能 と 複 雑 系 120－2
（2000．6．2）

相互依存関係にある内部モジュールを持つ自律システムの構築について

森山 甲一　沼尾 正行
東京工業大学 情報理工学研究科 計算工学専攻
koichi@nm.cs.titech.ac.jp, numao@cs.titech.ac.jp

概 要

複数の内部モジュールを持つ自律システムを構築する際に、モジュール間が完全な支配従属関係にあるとそのシステムの自律性について問題が生じる。そのため本研究では、支配従属ではなく相互に依存し合う内部モジュールを持つ自律システム（エージェント）を構築し、それをマルチエージェントの協調獲得問題に適用する。強化学習の分野でこの問題を解決する既存の手法を改良し、相互依存的な関係を持つ複数のモジュールとしてエージェントに実装した。このエージェントの集団と、監督者を有するエージェント集団などの異なるタイプのエージェントの集団とを比較する実験を行なった結果、特に動的環境では提案手法である相互依存的関係を持つエージェントが最も環境に適応することが分かった。

Construction of an Autonomous System with Interdependent Inner Modules

Koichi Moriyama and Masayuki Numao
Department of Computer Science,
Graduate School of Information Science and Engineering,
Tokyo Institute of Technology.
koichi@nm.cs.titech.ac.jp, numao@cs.titech.ac.jp

Abstract

When we construct an autonomous system by integrating several modules, there are troubles concerning the autonomy of the system if there is a module that dominates the whole system. Thus, we design the structure of a system (agent) having inner modules that are interdependent among themselves, and we apply the agents into a problem of obtaining cooperation of Multi-Agents. We enable a present method that can solve the problem in a reinforcement learning context to be applied into a dynamic environment, and the method is embodied into the agents as the interdependent modules. We conduct experiments comparing the proposed agents with agents such as those ones having a supervisor, and we confirm that the proposed agent having interdependent modules is the most flexible of all the tested agent.

1 はじめに

実行時に設計者の介入を要しない、真の意味での自律システムを構築することは重要であろう。実行時に設計者の介入を許さないのならば、設計者が予め、実行時に必要となる指針をシステムに導入しておけば良いのかもしれない。ところが、その指針がシステム全体を支配すると仮定するとシステムの自律性について問題が生じてしまう。その指針が直接的にシステムの挙動を左右するものであるならば、その挙動を環境から学習する機能をシステムに付与することによりその問題を回避することが出来る。ところが、例えば学習指針のように、その指針が間接的にしかシステムの挙動に影響しない場合にはどうすれば良いであろうか。そこで本研究では、上述の環境からの学習という考え方を拡張することにより、システム全体を支配する指針の代わりにその制御を受ける部分により影響を与えられて変化する指針を考える。この場合、制御する側とされる側の関係は支配従属ではなく相互依存的になる。本研究ではこの指針を1つの制御モジュールと見なし、内部にこれとその制御を受けるモジュールを導入したシステム（エージェント）を考える。そしてこの構造を持つ複数のエージェントを同一の環境で行動させ、その特性を調べることにする。

具体的には、まず同一環境にあるマルチエージェントの協調を強化学習の枠組で実現しようとする三上らの研究 [3, 4] を紹介する。それは強化学習に用いる強化信号（報酬）を予めフィルタリングするものであり、「共有地の悲劇」[1] 問題でそれなりの結果が得られている。そのため、このフィルタリング関数が一種の学習指針になっていると考えられる。しかし彼らのフィルタリング関数は固定的なものであるため、環境が動的に変化する場合に対応できない。そこで本研究では、この指針であるフィルタリング関数を、結果としてフィルタの制御を受けている強化学習プロセスからの情報により動的に変化させることを考える。そのように構築したエージェントの集団を、監督者の制御を受けるフィルタなどの異なるタイプのフィルタを持つエージェントの集団と比較するために、静的・動的環境における「共有地の悲劇」問題によって実験を行ない、結果を示す。

本論文は以下のように構成される。まず2節ではエージェントの概要を示し、エージェント全体を支配する指針を仮定することにより起こる問題を説明する。続いて3節では三上らの研究を紹介し、そのフィルタリング関数をパラメータ化する。更に、そのフィルタメカニズムを相互依存的な関係をもつモジュールとして2節で概要を示したエージェントに導入する。4節では、3節でフィルタメカニズムを導入したエージェントと、それとは異なるタイプのフィルタを持つエージェントについて「共有地の悲劇」問題でその性能を比較検証する。エージェント以外の環境要素が静的な場合と動的な場合の2種類の実験を行ない、その結果を示す。5節では4節の実験結果を考察し、6節で本論文をまとめ、今後の課題を述べる。

2 エージェントの概要

本研究で用いるエージェントは、学習部と制御部の2つのモジュールを持ち、外部情報を知覚部で知覚し、駆動部により外部環境に対して行動するものとする。通常我々がエージェントを構築する際に予め制御メカニズムを導入しようとする場合、その制御メカニズムがエージェント全体を支配するようにしてしまいがちである。このような支配的制御部を持つエージェントの概念図は 図1 となる。この図で実線矢印は制御を、点線矢印はデータの流れを表している。

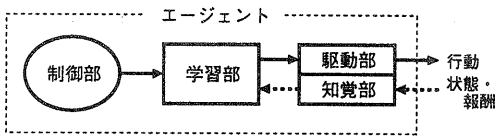

図 1 支配的な制御部を持つエージェント

制御部がエージェント全体を単に支配するだけであると問題が生じる。制御部が固定的な場合に

は、それはエージェントの全ての入力パターンに対して有効でなくてはならない。これはエージェントの存在する環境が非常に単純ならば可能であろうが、複雑な環境ではいわゆるフレーム問題やその類似問題 [5] を引き起こすと思われる。一方制御部が動的に変化するものである場合、その変化をどのように行なうかが問題となる。もしその変化に設計者などの人間がかかわる必要があるならば、もはやそのエージェントは自律的ではない。また、もしその変化のためのメタ制御部を導入するならば、同様の問題がそのメタ制御部に起こるだけである。

本研究ではこれらの問題を避けるために、このような支配的な制御部ではなく、制御を受けるモジュールからの影響を受ける相互依存的な制御部を持つエージェントを構築する。このようなエージェントの概念図は 図 2 となる。この図でアスタリスク (*) で示された矢印は、制御部と学習部の相互依存関係を表している。

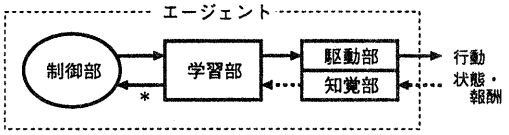

図 2　相互依存的な制御部を持つエージェント

3　エージェントの協調獲得

3.1　報酬のフィルタリング

マルチエージェントの協調獲得を強化学習の手法で行なうために、三上ら [3, 4] は強化学習プロセスにおける強化信号として報酬そのものではなく予めフィルタリングした報酬を用いることを提案した。この手法では報酬のフィルタリングに、自分を含む近隣のエージェントの平均報酬により計算される固定的なフィルタリング関数を用いている。文献 [3] では、マルチエージェントの問題のクラスに応じて平均フィルタと強調フィルタの 2 種のフィルタが定義されている。平均フィルタは報酬のうち近隣エージェントの平均報酬以上または以下の部分を平均報酬と同じ値にするものであり、強調フィルタは近隣エージェントの平均報酬以下の部分の値をより小さくすることでその部分を強調するものである。協調問題解決とデッドロック回避というマルチエージェントの相反する 2 つの大きなクラスに含まれる問題について、それぞれのフィルタが有効であることが文献 [3] に示されている。

3.2　汎用フィルタ

三上らによるフィルタは固定的なため、問題条件の変化する動的環境では有効でないと思われる。そこで本研究では以下の式で表されるパラメータ付きフィルタを提案する。

$$r' = \begin{cases} M + \alpha(r - M) & \text{if } r < M \\ M + \beta(r - M) & \text{otherwise.} \end{cases}$$

この式で r はエージェントの報酬、r' はエージェントのフィルタリング後の報酬（フィルタ後報酬）、M は近隣エージェントの平均報酬を表している。α, β はフィルタの性質を決定するパラメータである。このフィルタはパラメータを変化させることで文献 [3] に定義された全ての固定フィルタを包括するため、汎用フィルタと呼ぶことにする。

もしエージェントがこの汎用フィルタのパラメータを環境の変化に応じて適切に調節することが出来るならば、マルチエージェントの両方のクラスの問題それぞれについてこの汎用フィルタが有効に働くであろうし、環境が時間とともに変化する問題についても効果があるであろう。従って、エージェントが汎用フィルタのパラメータ制御方法を獲得することが重要になる。

3.3　エージェントへのフィルタの導入

ここでは、汎用フィルタを 2 節のエージェントの枠組みの中に導入することを考える。上述したように、汎用フィルタのパラメータ制御方法を獲得することは重要である。そこで本研究では、この

パラメータ制御方法を学習により獲得する手法を考える。**図2** におけるエージェントについて、パラメータ制御方法を学習により獲得する汎用フィルタを導入する際の変更点は以下の通りである。

- 学習部を強化学習部と汎用フィルタに分割
- 制御部をパラメータ学習部に変更

但し、この変更のみでは2つの問題が生じる。第一には **図2** におけるアスタリスクの付いた矢印の扱いである。この矢印をなくしてしまうと **図1** のようにエージェント内モジュールの関係は支配従属的になってしまう。第二にはパラメータの制御方法を学習するための手がかりとなるパラメータの評価に当たるものがないことである。そこで本研究では、パラメータ制御方法の学習に用いる評価信号をパラメータの影響を受ける強化学習部が与えることでエージェント内モジュールにおける相互依存関係を実現することにする。これらの変更を施したエージェントは **図3** となる。

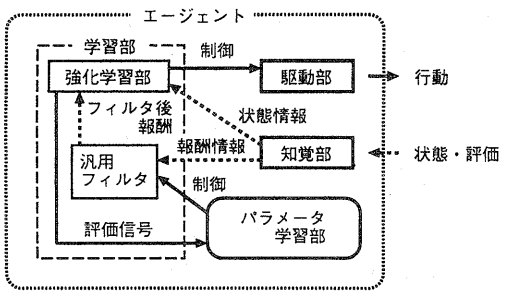

図3 汎用フィルタを導入したエージェント

本研究ではパラメータ学習部に与えられる評価信号を、強化学習部に与えられたフィルタ後報酬を符号反転したものとする。

4 実験

本節では、前節で汎用フィルタを導入したエージェントの性能を確認するため、**図3** のエージェントとそれとは異なるタイプのフィルタを持つエージェントを比較した実験を行なう。

4.1 準備

実験に用いる問題は、文献 [3] と同様に「共有地の悲劇」[1] タイプのゲームとする。「共有地の悲劇」とは、各プレイヤー（エージェント）が個人合理性に基づいて行動すると、ある共有財産の減少を引き起こし、結果として各プレイヤーの報酬が少なくなるという問題である。この問題では、より良い報酬を得るためには各プレイヤーが協力しなくてはならないため、協調問題解決の一種であると言える。

本研究で実験に用いるゲームでは、各エージェントは利己的・協力的・利他的のいずれかの行動を同時に提示し、その行動の組合せにより計算される報酬を獲得する。ここではこの一連の流れを1サイクルと称する。各エージェントに与えられる報酬は、他の条件が同じ時には利己的行動を提示した場合に最大を取り、利他的行動を提示した場合に最小となる。一方、利己的行動を提示した場合には負の共有財産である共有コストが増加し、利他的行動を提示した場合にはそれが減少する。従って全てのエージェントが利己的行動を提示した場合には、最終的に各エージェントの報酬は少なくなる。具体的には、一斉に行動を提示する際に0に初期化される共有コスト r_c が、あるエージェントが利己的行動を取った場合に Δr_c だけ増加し、利他的行動を取った場合には Δr_c だけ減少する。但し Δr_c は共有コスト r_c の変化パラメータである。共有コスト r_c をこのように全てのエージェントについて計算した後に各エージェントは報酬を獲得する。利己的・協力的・利他的行動を取ったエージェントに与えられる報酬はそれぞれ $3-r_c$, $1-r_c$, $-3-r_c$ である。

エージェントは合計10台であり、それぞれが固有の番号 i を持っている。エージェント a_i の近隣エージェント N_i を以下の式で定義する。N_i はエージェント a_i 自身を含むことに注意。

$$N_i \triangleq \{a_k \mid k = (i+j) \bmod 10, j = 0,1,2,3\}$$

実験に用いるエージェントが持つフィルタは大きく分けて次の4種類であり、そのうち汎用フィ

ルタはパラメータ学習（制御）方法により更に3つに分かれる。

- フィルタなし (NF)
- 完全平均フィルタ (FAF)

  文献 [3] で紹介されている2種の平均フィルタをまとめたものであり、エージェントの報酬を近隣エージェントの平均報酬と同一にする。

- 完全強調フィルタ (FEF)

  文献 [3] に紹介された強調フィルタについて、その相対する強調フィルタ[1]を構成してまとめたものであり、エージェントの報酬を近隣エージェントの平均報酬から離れるようにより大きくまたは小さくすることで強調するものである。文献 [3] における強調フィルタはパラメータを持つが、本研究ではそれを1に固定する。

- 汎用フィルタ

  3.2 節に述べたフィルタである。本研究では α, β それぞれのパラメータの制御を学習することにする。パラメータの初期値はフィルタなしと同等である $(\alpha, \beta) = (1, 1)$ とし、各パラメータの最小値を0、最大値を2とする。なおパラメータの変更はエージェントの行動後、強化学習部の学習前に行なうものとする。

  - 相互依存的評価 (GF-Int)

    3.3 節で述べたパラメータ評価法である。

  - 従属的評価 (GF-Sub)

    予め与えた固定期間の報酬合計をパラメータの評価とするものである。この期間をエージェントが変更できないことに注意。本研究ではこの期間を10サイクルとする。

  - 監督者による制御 (GF-Sv)

    全てのエージェントを監督するエージェントの存在を仮定し、その監督エージェントがその他全てのエージェントのパラメータを制御するものである。

[1] 3.1 節で紹介した通り、文献 [3] の強調フィルタは近隣エージェントの平均報酬以下の部分を強調するものである。そこで本研究ではそれに相対する強調フィルタとして、近隣エージェントの平均報酬以上の部分を強調するものを考える。

フィルタを構成する際に用いる近隣エージェントの平均報酬は、各エージェントにおけるフィルタ後報酬の平均とする。

強化学習部はもとより、汎用フィルタにおけるパラメータ学習部と監督エージェントの学習にも強化学習を用いる。手法としては Q-learning [6] を用い、学習率と割引率は共に 0.5 とする。学習結果からの行動選択は温度係数1のボルツマン分布 [2] による確率によって行なう。強化学習部における行動はエージェント自身の行動、状態は近隣エージェントの行動の組合せとし、強化信号はフィルタ後報酬とする。パラメータ学習部における行動はパラメータの変更（増加・維持・減少）、状態は1サイクル前に行なったパラメータの変更とし、強化信号は上述したパラメータ評価とする。監督エージェントの行動と状態はパラメータ学習部と等しく、強化信号は10台のエージェントの報酬合計とする。従属的評価による汎用フィルタのパラメータ変化頻度は評価期間と同じ10サイクルとし、その他の汎用フィルタのパラメータ変化頻度は1サイクルとする。その代わり、相互依存的評価と監督者による制御による汎用フィルタのパラメータ変化幅を0から0.1のランダム値とする一方、従属的評価による汎用フィルタのパラメータ変化幅を0から1のランダム値とする。

以下では、Δr_c を固定した静的環境における実験結果と、Δr_c をサイクル数に合わせて変化させた動的環境における実験結果を示す[2]。

4.2 静的環境における実験結果

図4に Δr_c を1に固定した静的環境における実験結果を示す。グラフの横軸はサイクル数、縦軸は10台のエージェントの報酬合計である。このグラフは4.1節で紹介したそれぞれのフィルタについて15回の実験結果の平均を5サイクルごとにプロットし、ベジエ近似したものである。

このグラフから以下のことが分かる。

[2] 静的環境とはエージェント以外の全ての要素が固定されている環境で、動的環境とはエージェント以外に変化する要素がある環境を言うものとする。

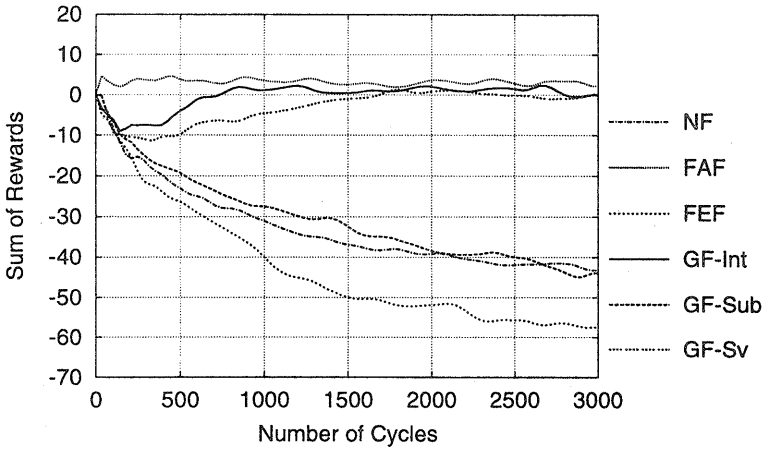

図 4　静的環境における実験結果

- フィルタなし (NF)・完全強調フィルタ (FEF)・汎用フィルタ（従属的評価 (GF-Sub)）の場合、サイクル数が大きくなるほど報酬合計が小さくなる。これは「共有地の悲劇」が起きていることを表している。

- 完全平均フィルタ (FAF) の場合、サイクル数に関係なくグラフがほぼ水平である。報酬合計が相対的に見て高いため、「共有地の悲劇」を回避することに成功していると見なせる。

- 汎用フィルタ（相互依存的評価 (GF-Int)・監督者による制御 (GF-Sv)）の場合、始めは報酬合計が減少するが途中で下げ止まり、上昇に転じて完全平均フィルタの結果に近付く。これはどちらの評価法も「共有地の悲劇」を回避するパラメータの学習が出来ていることを表している。

4.3　動的環境における実験結果

図 5 に Δr_c をサイクル数に応じて変化させた動的環境における実験結果を示す。グラフの横軸・縦軸と描画方法は静的環境と同様である。ここで Δr_c とサイクル数 c の関係は次式で表されるものとした。

$$\Delta r_c = \begin{cases} 1 - \dfrac{c}{1500} & \text{if } c < 1500 \\ 0 & \text{otherwise.} \end{cases}$$

この式はサイクル数が大きくなるほどエージェント間の協調が不要になることを意味している。

このグラフから以下のことが分かる。

- フィルタなし (NF)・完全強調フィルタ (FEF)・汎用フィルタ（従属的評価 (GF-Sub)）の場合、始めは報酬合計が減少し、最後には増加している。これは、フィルタの有効性が Δr_c の変化の影響を直接受けていることを表している。

- 完全平均フィルタ (FAF) の場合、やはりサイクル数に関係なくグラフがほぼ水平である。これは、このフィルタは協調が不要になるという環境の変化に適応できないことを表している。

- 汎用フィルタ（相互依存的評価 (GF-Int)・監督者による制御 (GF-Sv)）の場合、始めは完全平均フィルタに類似しようとするが、最終的にはフィルタなし・完全強調フィルタ・汎

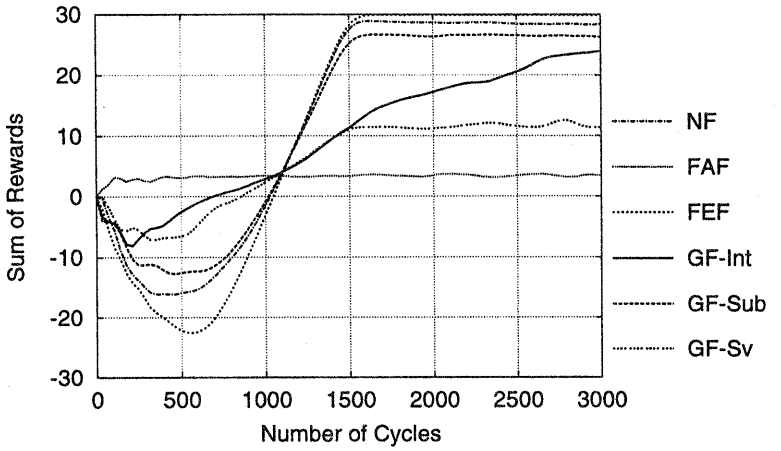

図 5　動的環境における実験結果

用フィルタ（従属的評価）に類似しようとする。これは、どちらの評価法も環境に応じてより良いフィルタのパラメータを学習することが出来ることを表している。興味深いことには、監督者による制御の場合は報酬合計が途中で頭打ちになるのに対して、相互依存的評価の場合は報酬合計が上昇し続ける。

5　考察

前節の実験では相互依存的評価を用いた汎用フィルタが最も環境に適応することが分かった。本研究では 3.3 節に述べたように相互依存的評価としてフィルタ後報酬を符号反転したものをパラメータ学習部に与えた。そこでまず、この相互依存的評価としてのフィルタ後報酬の符号反転について考察する。更により広い観点として、報酬をフィルタリングすることの社会的役割について簡単に考察することにする。

相互依存的評価としてのフィルタ後報酬の符号反転を考えるには、まずパラメータ学習部によって引き起こされる汎用フィルタのパラメータ変化の影響が実際のエージェントの行動に現れるまでの時間差を考慮に入れなくてはならない。汎用フィルタは駆動部ではなく学習部に直接の影響を与えるので、エージェントの行動に現れるまでには当然時間が掛かる。すなわちあるサイクルにおいてエージェントの行動へ影響を与えている汎用フィルタのパラメータは、そのサイクルで変更されたものではなく変更前のものである。そしてエージェントの獲得した報酬はこの行動によって得られたものである。よって、エージェントの獲得した報酬は変更前のパラメータの影響を強く受けていると考えられる。また、報酬が大きければ変更前のフィルタは環境に適応していると推測されるので、このサイクルで行なわれたパラメータの変更は不適切であったと考えられる。逆に報酬が小さければ、変更前のフィルタは環境に適応していないと推測されるので、パラメータの変更は適切であったと考えられる。更にエージェントが獲得した報酬とフィルタ後報酬が強く関係することについては明らかだろう。すなわち、フィルタ後報酬が小さい時にはパラメータの変更に大きな評価を与え、逆にフィルタ後報酬が大きい時にはパラメータの変更に小さな評価を与えることは理にかなうことである。本研究で用いたフィルタ後報酬の符号反転をパラメータの評価として与えることは、結果としてこれらの関係を表現しているのではないだ

ろうか。

続いて、報酬をフィルタリングすることの社会的役割について考察する。前節の実験では、完全平均フィルタを用いたエージェントの集合は、エージェント間の協調が必要とされる状況の下では良い結果を残している。ところで、フィルタがエージェントの報酬を操作するのはエージェントが自分の行動を学習する前であり、その操作はエージェントの内部で行なわれる。監督者によって制御される汎用フィルタを持つエージェントでもこれは同じである。つまり、報酬がフィルタによって制御されるのはエージェントの外ではなく内部であるため、このフィルタというものはエージェントが自分の獲得した報酬をどのように見なすかという一種の感情的・性格的機構であると思われる。このように考えると、完全平均フィルタを持つエージェントは性格的に報酬を追求しないように設計されたエージェントであると言える。この場合、このエージェントはゲーム理論的にもはや合理的ではないため、合理的エージェントの集団が陥る「共有地の悲劇」を回避することが出来るのである。

6 まとめ

本研究では相互に依存する2つの内部モジュールを持ったエージェントを構築し、それをマルチエージェントの協調獲得問題に適用した。その問題を解くための手段として三上らによる報酬のフィルタリングを用いた手法を紹介し、それに用いられた固定フィルタをパラメータ化することで環境に動的に適応することの出来る汎用フィルタを提案した。汎用フィルタの適切なパラメータを学習によって獲得することとし、その学習に必要なパラメータ評価を、汎用フィルタの制御を受ける強化学習部が与えることによって相互依存的な評価を行なうこととした。この評価としては具体的にフィルタ後報酬の符号反転を与えることとした。この相互依存的評価による汎用フィルタの性能を、マルチエージェント全体を監督する監督者の制御を受ける汎用フィルタなどの他のフィルタの性能と比較するために、静的・動的環境において「共有地の悲劇」タイプのゲームによる実験が行なわれた。その結果、特に動的環境においては提案手法である相互依存的評価による汎用フィルタが環境に最も適応することが示された。更に、この実験ではどうして相互依存的評価としてのフィルタ後報酬の符号反転が良い結果をもたらしたのかということについて、また報酬をフィルタリングすることの社会的役割について考察を行なった。

今後の課題としては、まず本研究で用いた相互依存的評価、つまりフィルタ後報酬の符号反転の数学的性質を明らかにすることが挙げられる。これには、エージェントの設計に関するシステムの理論やマルチエージェントの数学的解析を行なって来たゲーム理論などの分野における知見が利用できるのではないかと思われる。更にこの符号反転という操作が2節で否定した全体を支配する一種の指針となっている点が否めないため、この符号反転という操作をすることなく相互依存的関係を導く必要があると思われる。これは、エージェント内の学習部が自動的に制御部の評価方法を獲得することを意味している。

References

[1] Garrett Hardin. The Tragedy of the Commons. *Science*, Vol. 162, pp. 1243–1248, 1968.

[2] Leslie P. Kaelbling, Michael L. Littman, and Andrew W. Moore. Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, Vol. 4, pp. 237–285, 1996.

[3] Sadayoshi Mikami and Yukinori Kakazu. Cooperation of Multiple Agents Through Filtering Payoff. In *1st European Workshop for Reinforcement Learning*, 1994.

[4] Sadayoshi Mikami, Yukinori Kakazu, and Terence C. Fogarty. Co-operative Reinforcement Learning By Payoff Filters. In *Proc. 8th European Conference on Machine Learning, ECML-95*, (Lecture Notes in Artificial Intelligence 912), pp. 319–322, Heraclion, Crete, Greece, 1995.

[5] Stuart J. Russell and Peter Norwig. *Artificial Intelligence: A Modern Approach.* Englewood Cliffs, NJ: Prentice-Hall, 1995.

[6] Christopher J. C. H. Watkins and Peter Dayan. Technical Note: Q-learning. *Machine Learning*, Vol. 8, pp. 279–292, 1992.