様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　６月　３日現在

機関番号：１４３０３
研究種目：基盤研究(C)
研究期間：2010～2012
課題番号：22500131
研究課題名（和文）　複雑な学習問題に対するＰＳＯに基づく群強化学習法の適用
研究課題名（英文）　Swarm Reinforcement Learning Methods Based on PSO for Complicated Learning Problems

研究代表者
飯間　等（IIMA　HITOSHI）
京都工芸繊維大学・工芸科学研究科・准教授
研究者番号：70273547

研究成果の概要（和文）：短時間で最適な方策を学習するために、Particle Swarm Optimization(PSO)に基づく群強化学習法を提案し、連続状態行動空間を有する問題などの複雑な強化学習問題に提案方法を適用した。提案方法はエージェントと環境の組（これを学習世界と呼ぶ）を複数用意し、各学習世界のエージェントが個別に通常の強化学習法を用いて学習を行うとともに、PSO の更新式を用いた学習世界間の情報交換による学習も行う方法である。

研究成果の概要（英文）：We proposed swarm reinforcement learning methods based on particle swarm optimization (PSO) for acquiring optimal policies rapidly, and applied the proposed methods to some complicated reinforcement learning problems such as ones with continuous state-action space. In the proposed method, multiple sets of an agent and an environment, which are called learning worlds, are prepared, and agents in each learning world learn not only by individually using a usual reinforcement learning method but also through exchanging information among the learning worlds by using the update equations of PSO.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	2,000,000	600,000	2,600,000
２０１１年度	800,000	240,000	1,040,000
２０１２年度	500,000	150,000	650,000
年度			
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：知能情報学
科研費の分科・細目：情報学・知能情報学
キーワード：(1) 強化学習 (2) PSO (3) 群知能

１．研究開始当初の背景

強化学習とは、エージェントが試行錯誤を通して自身の目的を達成するための最適な行動選択方針（これを方策という）を獲得しようとする機械学習の一種である。エージェントは現在の状態を観測し、何らかの行動を行うと別の状態に遷移し、報酬を受け取る。このとき、報酬の総和を最大にするための方策を決定する問題が強化学習問題である。

従来の強化学習法はエージェントが試行錯誤を通して学習を行うため、問題が複雑になると最適な方策を得るまでに膨大な学習時間が必要となるという欠点を有しており、実用化に向けて学習を高速に行う新しい方法を開発することが急務となっている。ここで、最適化の分野に目を向けると、遺伝的アルゴリズムや Particle Swarm Optimization (PSO)のように複数の個体（エージェント）

を用いた多点探索最適化法が、複雑な問題に対して短時間に最適解を発見できる能力を有していることが知られている。

そこで、エージェントと環境の組（これを学習世界と呼ぶ）を複数用意して学習を行うことにより短時間で学習できると考え、この考え方に基づく強化学習法を以前に提案し、それを群強化学習法と呼んでいる。群強化学習法は各学習世界で個別に通常の強化学習法を実行して学習を行うとともに、個別学習で優れた学習を行った学習世界の何らかの情報を学習世界間で交換することによっても学習を行う方法である。平成 19 年度から萌芽研究「群強化学習法の開発」を遂行し、主として状態行動空間が離散値の基本的な問題を対象として学習世界間の情報交換方法を検討した。その結果 PS0 の解候補更新式を利用して情報交換を行う群強化学習法が非常に短時間で学習できることを突きとめた。そこで、この方法をさらに発展させることにより、複雑で現実的な強化学習問題に対する最適方策を短時間に獲得する方法を開発することができると考えた。

２．研究の目的

本研究の全体構想は、複数の学習世界を用意して学習を行う群強化学習法を種々の問題に対して開発することである。その中で、本研究の具体的な目的は、短時間に最適な方策を獲得するための PS0 に基づく群強化学習法を提案し、提案方法を現実に存在するような複雑な強化学習問題に適用することである。群強化学習法の性能は学習世界間の情報交換法に強く依存する。そこで、まず PS0 の更新式を用いた学習世界間の情報交換法を開発する。つぎに、現実に存在するような複雑な強化学習問題をいくつか取り上げ、これらの問題に PS0 に基づく群強化学習法を適用する。複雑な問題にはそれぞれ固有の難しさがあるので、各問題向けの群強化学習法を開発する。

３．研究の方法

(1) 複雑な強化学習問題の設定

本研究では現実に存在するような複雑な強化学習問題に対して PS0 に基づく群強化学習法を適用するが、強化学習問題は多種多様であり、それゆえある程度問題の種類ごとに群強化学習法を開発する必要がある。そこで、ここでは現実に存在するような複雑な問題を三つ設定する。一つ目の問題として、状態行動空間が連続である問題を扱う。このような問題はロボットなどの制御器を設計する問題などにしばしば現れる。二つ目の問題として、エージェントが複数存在するマルチエージェント問題、この中でも特に近年注目を集めているフォーメーション制御に関する問題を扱う。最後に、学習の目的が一つではなく複数存在する多目的の問題を扱う。従来から研究されている多くの強化学習問題では単純化した問題を扱うために目的を一つとしているが、実際の問題では複数の目的を有することが多いので、このような多目的問題を扱う。

(2) 以前に提案した群強化学習法に関する問題点の整理

以前に提案した PS0 に基づく群強化学習法は短時間に良い方策を発見できるものの、その方策は必ずしも最適ではない。そこで、以前に提案した群強化学習法で最適方策が得られない問題点を整理する。また、以前に提案した群強化学習法は状態行動空間が離散値の基本的な問題に適用しているだけで、複雑な問題には適用していなかった。 したがって、(1)で設定した複雑な問題に群強化学習法を適用する上で、何らかの問題が発生する可能性があるので、ここでそのような問題点を整理する。

(3) 複雑な問題に対する群強化学習法の開発

最適方策が得られないという群強化学習法の問題点を解決する方法およびアルゴリズムを開発する。また、複雑な問題に群強化学習法を適用する上で発生する問題点を解決する方法およびアルゴリズムを開発する。複雑な問題に対しては、(1)で設定した連続状態行動空間の問題、マルチエージェント問題、多目的問題のそれぞれの問題ごとに、方法およびアルゴリズムを開発する。

(4) 開発した群強化学習法の性能評価

(1)で設定した三つの複雑な問題の例題を用意し、これらの例題に(3)で開発した群強化学習法を適用することで、群強化学習法の性能を評価する。連続状態行動空間の例題としてロボットの歩行制御器を設計する問題、マルチエージェントの例題として複数台のロボットでフォーメーションを形成する問題、多目的の例題として深海にある宝を探索する問題を用意する。宝探索問題の目的は二つあり、一つは価値の高い宝を見つけることで、もう一つは早く宝を見つけることである。用意した例題に対して、開発した PS0 に基づく群強化学習法、他の情報交換法を用いる群強化学習法、および他の強化学習法を適用し、これらの実行結果を比較することで、開発した群強化学習法の性能を評価する。

４．研究成果

(1) 寿命を設定した自己最良値を用いる PS0 に基づく群強化学習法の提案

以前に提案した PS0 に基づく群強化学習法では、各学習世界で個別に実行する通常の強化学習法による学習と、PS0 の更新式を用いた学習世界間の情報交換による学習が交互

に行われる。後者の学習では、各学習世界がこれまでに発見した状態行動価値（Q 値）の中で最も優れたQ値（自己最良Q値と呼ぶ）と、全学習世界の自己最良Q値の中で最も優れたQ値（全体最良Q値と呼ぶ）を用いて学習を行う。これらの最良Q値を選別するために、個別の学習を終えた後に各学習世界の Q 値を評価する。ここで、Q 値の評価は近似的に行っており、それゆえ必ずしも適切に評価が行われるとは限らず、Q 値が過大評価されることがある。この以前に提案した群強化学習法では最適方策が得られなかったが、それはこのような過大評価された最良Q値を用いて学習を行ったためであると考えられる。

そこで本研究では、短時間に最適方策を得るために、自己最良Q値に寿命を設定する方法を提案した。提案方法では、寿命分のエピソード数が経過すると、過去のある一定エピソード数内で優れていると評価されたQ値で自己最良Q値を置き換える。また、全体最良Q値がこの置き換えられた自己最良Q値となっている場合は、全体最良Q値も選別し直して置き換える。この方法により、たとえ過大評価されたQ値が最良Q値となったとしても、それが寿命を迎えた時点で別のQ値に置き換えられるので、短時間に最適方策を獲得できることが期待できる。

提案方法の性能を評価するために、二次元格子平面上のスタート座標からゴール座標まで最少の行動回数で到達する方策を学習する問題に提案方法および他の学習法を適用し、それらの結果を比較する数値実験を行った。各エピソードにおけるエージェントの行動回数の推移を図１に示す。この図の横軸はエピソード数を示し、群強化学習法の場合は全学習世界のエピソード数の総和で示されている。図１より、Q-learning 法より群強化学習法の方が、少ないエピソード数で行動回数が大きく減少していることがわかる。また、寿命を設定しない群強化学習法より寿命

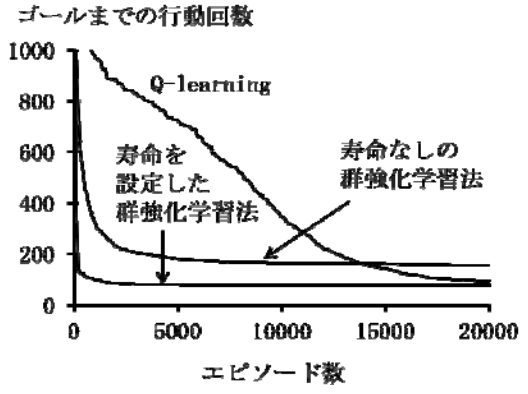

図１　最短経路問題に各学習法を適用したときの各エピソードにおけるエージェントの行動回数の推移

を設定した提案方法の方が行動回数は少なく、提案方法の有効性を確認できる。

(2) 連続状態行動空間の問題に対する群強化学習法の適用

以前に提案した群強化学習の個別学習法では連続状態行動空間の問題を扱うことはできない。そこで連続の問題に適用するために、連続の問題に対する代表的な強化学習法として知られている Actor-Critic 法を個別学習に用いる群強化学習法を提案した。Actor-Critic 法は方策を学習する Actor と価値関数を学習する Critic の２つの学習器を用いる方法である。提案方法では、優れていると評価された学習世界の方策と価値関数を学習世界間で交換して学習を行う。提案した方法を倒立振子制御系設計問題に適用し、PSO に基づく情報交換法を用いる提案方法によって短時間に振子を倒立させる制御系を設計できることを確認した。

Critic で価値関数を学習する方法として、基底関数に基づく関数近似法がしばしば用いられる。特に、価値関数が高次元で非線形性が強い場合には、基底関数の個数、中心座標、分散および係数で価値関数をパラメトライズしておき、これらのパラメータの値を適応的に変化させる方法が用いられる。ところが、この方法を群強化学習法の個別学習で用いると、各学習世界でパラメータの個数が異なるために以前に提案した PSO に基づく学習世界間の情報交換法を適用することができなくなるという問題が生じる。

そこで本研究では、価値関数を PSO の更新式で更新し、更新後の価値関数における学習パラメータの値を求めることで、学習パラメータの値を更新する。ところが、PSO の更新式適用後の価値関数（これを目標関数と呼ぶ）における学習パラメータの値を求めようとすると、実質的には最初から関数近似問題を解き直すことに相当し、これを行うには莫大な計算時間を必要としてしまう。この問題を解決するために、価値関数を目標関数そのものに更新するのではなく、目標関数との差のノルムが少ない関数に短時間で更新する方法を提案する。この提案方法では、各学習世界におけるエージェントの基底関数はそのまま用いる。すなわち、基底関数の個数、中心座標および分散は更新しない。その上で、各基底関数の中心座標における関数値が、同じ座標における目標関数値に一致するように基底関数の係数のみを更新する。この方法によって中心座標の近傍において更新後の価値関数と目標関数との差のノルムが減少し、全体のノルムも減少することが期待できる。

提案方法の性能を評価するために、二足ロボットを５秒間歩行させる制御系を設計する問題に提案方法および他の学習法を適用

し、それらの結果を比較する数値実験を行った。各学習法で獲得した方策を用いてロボットを歩行させるシミュレーションを行ったときの歩行時間を表１に示す。表１において、BEST と AVE は群強化学習法における他の情報交換法の名前である。BEST は最も優れていると評価された学習パラメータの値を他の学習世界にコピーする方法であり、AVE はその最も優れていると評価された値と各学習世界での値の平均値をとる方法である。表１より、PSO を用いる提案方法によって歩行時間の長いロボットを実現できていることがわかる。

表１　二足歩行ロボット制御系設計問題に各学習法を適用して得られた方策を用いたシミュレーションにおけるロボットの歩行時間（秒）

方法	群強化学習法			Actor-Critic 法
	BEST	AVE	PSO	
歩行時間	3.403	4.532	4.831	0.637

価値関数を少ないメモリで近似する方法としてタイル型の関数を組み合わせる方法が用いられる。そして、タイル型関数の個数、サイズおよび位置を、複数の個体を用意して適応的に学習する方法として進化的タイルコーディング法が提案されている。この方法では二分木を用いてタイル型関数を表現して、この二分木を適応的に変化させる。ところが、個体間で情報が交換されておらず、それゆえ多点探索法の利点が活かされていない。そこで、個体間で情報を交換する PSO を用いて各タイル型関数の個数などを適応的に変化させて学習を行う方法を提案した。提案方法では、進化的タイルコーディング法と同様に二分木を用いてタイル型関数を表現し、ノードごとに PSO の更新式を適用する。ところが、各学習世界で二分木の形が異なると PSO の更新式が適用できないという問題が生じる。そこで本研究では、仮想ノードを追加するとともに深さを全学習世界で統一した完全二分木によってタイル型関数を表現する方法を提案した。提案した方法を倒立振子制御系設計問題に適用し、提案方法によって短時間に振子を倒立させる制御系を設計できることを確認した。

(3) マルチエージェント問題に対する群強化学習法の適用

マルチエージェント問題はエージェント同士の関係などにより様々な種類の問題があるため、ある程度問題の種類毎に学習法を開発する必要がある。本研究では、複数のロボットが何らかのフォーメーションを形成するために、各自の初期座標から目標座標までの最短経路を学習する問題、より正確には全ロボットの行動回数の総和が最小となる方策を学習する問題を扱う。このような問題は人文字のようにロボットが集まって文字や絵を描くような場合に現れる。目標座標はロボットの台数と同じ個数だけ存在し、各ロボットは互いに異なる目標座標に到達しなければならない。どのロボットがどの目標座標に向かうかはあらかじめ与えられず、それを学習によって獲得させる。

このフォーメーション形成問題に対して、各ロボットが適切な目標座標に向かっているかどうかを考慮しながら各学習世界の Q 値を評価して学習世界間で情報交換を行う群強化学習法を提案した。Q 値の評価方法として、各ロボットの行動回数を用いるのが一見妥当に思える。ところが、行動回数のみで評価すると、目標座標に近い初期座標から移動するロボットがその目標座標へ向かう方策を学習する可能性がある。その結果、他のロボットがより遠くの目標座標に移動することになり、このような方策はロボット全体としては最適でない可能性が高い。一方、全てのロボットが最適な目標座標に最少回数の行動で到達できている状況を考えると、このときの各ロボットの行動回数のばらつきは小さいと考えられる。そこで提案方法では、全ロボットの行動回数の総和と分散に基づいて Q 値を評価し、この評価結果に基づいて選別された優れた Q 値を学習世界間で交換する。

提案方法の性能を評価するために、提案方法および他の学習法を例題に適用し、それらの結果を比較する数値実験を行った。各エピソードにおける全エージェントの総行動回数の推移を図２に示す。これより、約 550 エピソード以降では PSO を用いる提案方法における総行動回数が最も少なくなっており、提案方法の有効性が確認できる。

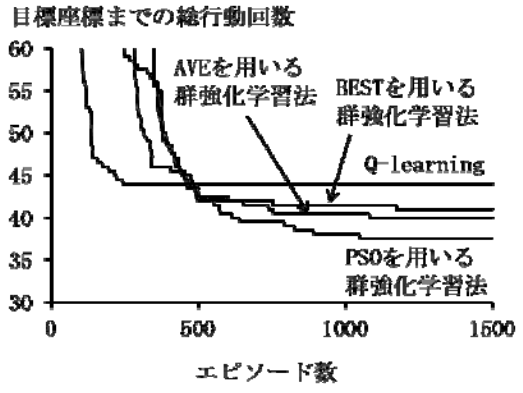

図２　フォーメーション形成問題に各学習法を適用したときの各エピソードにおける全エージェントの総行動回数の推移

(4) 多目的問題に対する群強化学習法の基礎的検討

複数の目的に対応した複数種類の報酬をエージェントが受け取る多目的問題を考える。二つの方策 A,B に対して、方策 A にしたがって行動したときに得られる報酬の総和がどの種類においても方策 B にしたがって行動したときに得られる報酬の総和以上で、かつ一つ以上の種類に対しては方策 A による報酬の総和が方策 B による報酬の総和より大きくなるとき、方策 A は方策 B を優越するという。このとき、他のどの方策にも優越されない方策が多目的問題における最適な方策であり、この方策をパレート最適方策と呼ぶ。パレート最適方策は一般に数多く存在する。

以前に提案した群強化学習法は多目的問題を扱うことはできない。そこで、多目的問題における全てのパレート最適方策を求める群強化学習法の開発を試みた。単目的の問題に対する群強化学習法の個別学習では、通常は以前に提案されている代表的な強化学習法を用いればよい。ところが、多目的問題に対する強化学習法の研究はまだ十分に行われていない。そこで、群強化学習法の個別学習に用いる強化学習法の開発を目指して、全てのパレート最適方策を発見する強化学習法を開発する。次状態への遷移確率や報酬の期待値が既知の問題に対しては、Q ベクトルの凸包の頂点を用いた価値反復法が提案されている。そこでこの方法の考え方を用いて、次状態への遷移確率などが未知である多目的強化学習問題に対する学習法を提案した。

提案方法では、単目的問題で用いられる Q-learning 法の更新式における Q 値を Q ベクトルの凸包の頂点集合（Q 凸包集合と呼ぶ）に置き換えて、この Q 凸包集合を更新することで学習を行う。ところが、この方法ではパレート最適方策への収束に寄与しない多くの Q ベクトルが Q 凸包集合に含まれており、その無駄な Q ベクトルの更新に多大な時間を必要とする。そこで、より短時間に学習する方法として、Q 凸包集合の代わりに優越されない Q ベクトルの集合を更新して学習する方法も提案した。

提案した二つの方法によって得られる方策がパレート最適方策になることを理論的に証明した。また、提案方法の性能を評価するために、エージェントが深海にある宝を発見することを学習する問題に提案方法を適用する数値実験を行った。この問題の目的は二つあり、一つは価値の高い宝を見つけることで、もう一つは早く宝を見つけることである。この問題にはパレート最適方策が１０個存在するが、いずれの提案方法によっても全てのパレート最適方策を発見することができた。また、各提案方法の計算時間を表２に示す。これより、優越関係を用いる提案方法の方が凸包を用いる提案方法よりも計算時間が短いことがわかる。ただし、優越関係を用いる提案方法でも多くの計算時間を必要とする。この提案方法を個別学習に用いた PSO に基づく群強化学習法を開発することにより、短時間に学習することが期待できる。

表２　宝探索多目的問題における各提案方法の計算時間（秒）

方法	計算時間
凸包を用いる方法	689176
優越関係を用いる方法	1605

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計８件）

① 飯間　等、黒江康明、連続状態行動空間を有する問題に対する群強化学習法、計測自動制御学会論文集、査読有、48 巻、2012、790-798
DOI: 10.9746/sicetr.48.790

② Yusuke Mukai、Yasuaki Kuroe、Hitoshi Iima、Multi-Objective Reinforcement Learning Method for Acquiring All Pareto Optimal Policies Simultaneously、Proceedings of 2012 IEEE International Conference on Systems, Man and Cybernetics、査読有、2012、1917-1923
DOI: 10.1109/ICSMC.2012.6378018

③ Hitoshi Iima、Yasuaki Kuroe、Kazuo Emoto、Swarm Reinforcement Learning Methods for Problems with Continuous State-Action Space、Proceedings of 2011 IEEE International Conference on Systems, Man and Cybernetics、査読有、2011、2173-2180
DOI: 10.1109/ICSMC.2011.6083999

④ Hitoshi Iima、Yasuaki Kuroe、Swarm Reinforcement Learning Method Based on an Actor-Critic、Proceedings of Eighth International Conference on Simulated Evolution and Learning、査読有、2010
DOI: 10.1007/978-3-642-17298-4_29

〔学会発表〕（計１４件）

① 飯間　等、フォーメーション形成問題に対する Particle Swarm Optimization に基づく群強化学習法、第 57 回システム制御情報学会研究発表講演会、2013 年 5 月 17 日、兵庫県民会館

② 伊藤　洋、Particle Swarm Optimization に基づくタイルコーディングを用いた強化学習法、計測自動制御学会第 40 回知能

システムシンポジウム、2013 年 3 月 14 日、京都工芸繊維大学

③ 飯間　等、高次元連続状態行動空間の問題に対する群強化学習法、計測自動制御学会システム・情報部門学術講演会 2011、2011 年 11 月 21 日、東京都国立オリンピック記念青少年総合センター

④ 飯間　等、寿命を設定した自己最良値を用いた Particle Swarm Optimization に基づく群強化学習法、計測自動制御学会システム・情報部門学術講演会 2010 講演論文集、2010 年 11 月 25 日、キャンパスプラザ京都

⑤ 飯間　等、寿命のある自己最良値を用いた Particle Swarm Optimization に基づく群強化学習法、第 54 回システム制御情報学会研究発表講演会、2010 年 5 月 19 日、京都リサーチパーク

６．研究組織

(1)研究代表者

飯間　等（IIMA　HITOSHI）

京都工芸繊維大学・工芸科学研究科・准教授

研究者番号：70273547

(2)研究分担者

黒江　康明（KUROE　YASUAKI）

京都工芸繊維大学・工芸科学研究科・教授

研究者番号：10153397

(3)連携研究者

（　　）

研究者番号：