様式 C-19

科学研究費補助金研究成果報告書

平成２２年３月３１日現在

研究種目：若手研究（Ｂ）
研究期間：２００７～２００９
課題番号：１９７４００６０
研究課題名（和文）意思決定過程における時間差分制御の研究とその応用

研究課題名（英文）On study of temporal difference method in decision process and its application

研究代表者
堀口　正之（HORIGUCHI MASAYUKI）
研究者番号：９０３６６４０１

研究成果の概要（和文）：本研究は、不確実な環境下での意思決定過程において従来研究されてきた動的計画法による理論的な最適性について、意思決定者の状態観測と行動決定の学習に基づく評価関数の推定による実用上の計算困難性の克服を目的としている。未知の推移法則をもつ数理モデルにおける学習アルゴリズムの理論構築とその応用研究を、ニューロダイナミックプログラミングによる時間差分法と区間ベイズ推定法によるモデル推定から最適解の探索手法を明らかにし、シミュレーションによる理論の数値実験、学習アルゴリズムの改良に取り組んだ。

研究成果の概要（英文）：In decision process with uncertainty, the optimal solutions are constructed by using dynamic programming(DP) algorithms. In order to solve practical problems involving very large state space, we need to decrease the amount of computation necessary for learning algorithm since DP algorithm cannot be applied directly. Based on using the data of state-action process, the value function is estimated by learning algorithm. We consider temporal difference method of Neuro Dynamic Programming (Neuro-DP) and Bayesian interval estimation in Markov decision processes with unknown transition law. We derive algorithms of constructing optimal solution theoretically. We also treat numerical examples to show the validity of algorithms and improve corresponding algorithms.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	１，２００，０００	０	１，２００，０００
２００８年度	６００，０００	１８０，０００	７８０，０００
２００９年度	６００，０００	１８０，０００	７８０，０００
年度			
年度			
総　計	２，４００，０００	３６０，０００	２，７６０，０００

研究分野：計画数学
科研費の分科・細目：数学・数学一般（含確率論・統計数学）
キーワード：マルコフ決定過程、計画数学、適応政策、学習理論、マルコフ集合連鎖

１．研究開始当初の背景

意思決定過程の代表的モデルの一つであるマルコフ決定過程(MDP)において最適解を導くアルゴリズムとして Howard の政策反復法(PIM)や、さらに状態数が比較的大きい場合に対応した修正政策反復法が多く研究されている。しかしながら、これらの手法については、いわゆる次元の呪いの隘路の問題が存在し、状態数が増えるにつれて計算量が爆発的に増大して実用規模の問題を解くことが困難であることが知られている。他方、最適解の近似解法として Value Iteration Method(VIM)が知られている。しかし、PIM や VIM では状態推移が一つの閉じた系内において考察されることがほとんどで、多重マルコフ連鎖(multi-chain)での VIM の研究は未解決の問題が多くあり、モデルを構成する要素、特に、状態間での推移の仕方によって計算手順の進捗が影響される。

不確実な環境を含む意思決定過程での数理モデルとして推移法則が未知の場合を扱い、そのモデルにおける最適政策を導く学習アルゴリズムの理論構築と実用的な解析手法の研究が緊急の課題である。

２．研究の目的

意思決定過程において実用規模の問題では計算困難とされる数理モデルに対する Neuro-DP(ニューロ・ダイナミック・プログラミング)の適用可能性について、モデルを構成するパラメータの変化にも柔軟に対応できる頑健性ある最適化アルゴリズムを開発することを目的とする。

具体的には、次の３つの課題を中心に研究を進める。

(1) 不完全情報の状況下でのマルコフ決定過程における学習理論

(2) Neuro-DP による学習アルゴリズムの数理の研究

(3) 時間差分的近似アルゴリズムの実際的問題への応用

３．研究の方法

(1)マルコフ決定過程における学習理論の研究として、Neuro-DP による時間差分的近似法について PIM や VIM の先行研究の成果と比較しながら収束の速さや誤差範囲、統計的推測に基づく多段意思決定の取り方など計算アルゴリズムの特性についてその数学理論を研究する。

(2)communicating class を持つ場合と多重連鎖を持つ場合について、時間差分法、Q-learning、Actor-Critic Method などの概念に基づいた強化学習の適用可能性を考察する。

(3)アルゴリズムの計算プログラム作成によりコンピュータシミュレーションを実行し、アルゴリズム実行による逐次近似解の挙動分析を行い、アルゴリズム改良に取り組む。

(4) 凸解析や非線形計画法の視点からの新たな学習アルゴリズムの開発や停止決定過程における学習理論の枠組みについて研究する。

(5)意思決定過程における学習理論の研究として、不確実な状況下でのベイズ理論に基づく最適化手法について、その数学理論を研究する。

４．研究成果

(1)有限個の状態数を持つ推移法則未知のマルコフ決定過程において、時間差分法(Temporal Difference Method)による最適な適応政策の存在と学習アルゴリズムの研究を行った。具体的には、推移法則の集合族について、

①すべての状態間に互いに１期間で推移できる正の確率を持つ場合

②状態集合の、ある部分集合に属する任意の２つの状態間が互いに到達可能であり(communicating class)、それ以外の補集合の状態はすべて過渡的状態(transient class)である場合

について考察した。①では、推移法則の推定に履歴による最尤推定を用い、時間差分による適応型決定の取り方として修正 greedy policy を導入して適応政策の最適性を明らかにした。②の場合では、先行研究で得ているマルコフ連鎖の推移状況から推測される状態集合の構造を学習するアルゴリズムを適用し、割引利得最適化問題からの近似理論とgreedy policyを取る学習アルゴリズムにより、最適な適応政策が構成できることを明らかにした。また、その学習アルゴリズムの数値シミュレーションも行い、アルゴリズムの有効性を明らかにした。

本研究成果により、不完全な情報をもつ２つの意思決定モデルでの適応型最適政策の構成方法や学習アルゴリズムの数理とその有効性を明らかにした。

(2)不確実性下でのマルコフ決定過程における学習理論として、ベイズ推定を用いた最適化手法の構築に取り組んだ。

先行研究では、推移回数の過去の履歴から構成される相対頻度による最尤推定法を

用いた場合やベイズの定理に基づく事前分布の更新が用いられている。新たな手法として、事前知識としての区間測度からベイズ理論による推移法則の区間推定を取り入れた数理モデル（区間ベイズマルコフ決定過程）を構成した。事前分布として、確率分布でない区間で表現された測度を導入することで、推移法則や価値関数(Value Function)も事後解析では実数値の閉区間として表現することができる。また、事後知識の解釈や活用について、事前分布を必ずしも一つの分布として仮定しない場合でもモデルの事前・事後解析が可能であり、先行研究のものよりも実用的なものとなったと言える。

さらに、この新たなマルコフ決定過程において、状態推移の履歴から推定された区間推移確率行列の連続性や収束性、価値関数の収束性などの性質を導いた。また、具体的な数値実験によりこのベイズ推定手法について例示した。推移法則が区間表現される意思決定モデルでの政策に関する学習理論を、今後さらに考察するための重要な成果の一つとなった。

また、時間差分制御や適応学習理論に関するこれまでの研究成果について、ひとつの論文として総合的に取りまとめた。

(3)マルコフ決定過程における未知の推移法則の推定方法について、状態の逐次観測から得られるデータセットに基づいた区間ベイズ推定法を用いて、推移確率行列の区間表現を得るための計算手法の考察を行った。

推定される推移確率行列は、それぞれの成分が事後区間として表され、データ観測数に依存した高次多項方程式の解として特徴づけられる。ニュートンラフソン法と不完全ベータ関数により、その数値近似解を得る計算アルゴリズムを示した。また、不完全ベータ関数によって表される高次方程式の解法と分数計画問題における双対問題の解法との関係についても明らかにした。これらの手法を元にして、観測から得る任意のデータセットに対して区間推移確率行列を容易に得ることが出来る。これは、実際問題への応用について考察する際の有効な計算手法を明らかにしたと言える。

また、事後区間の推定方法として、確信区間に基づく区間表現方法についても考察を行った。具体的には、事後測度区間によるパーセンタイルを求め、確信度に応じた推移確率行列の事後区間表現について区間の上限および下限を不完全ベータ関数によって表現できることも示した。

これらは、推移確率行列が区間で表現されるマルコフ決定過程において、状態観測から事後区間を推定する逐次学習を行っていく新たな適応型学習の手法に貢献できるものと考えられる。

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９件）

① 伊喜哲一郎、堀口正之、安田正實、蔵野正美、“不確実性の下でのマルコフ決定過程に対する区間ベイズ手法”、京都大学数理解析研究所講究録 1636「不確実性と意思決定の数理」、査読無、2009, 1--8

② 岩村 覚三、堀口 正之、堀池 真琴、“ダイナミックプログラミングを用いたファジィメトリッククラスタリング(Fuzzy Metric Clustering Based on Dynamic Programming)”、京都大学数理解析研究所講究録 1630「非加法性の数理と情報：非加法性と凸解析」、査読無、2009, 77--88

③ 伊喜哲一郎、堀口正之、蔵野正美、安田正實、“A pattern-matrix learning algorithm for adaptive MDPs: The regularly communicating case”、京都大学数理解析研究所講究録 1589「不確実な状況における意思決定の理論と応用」、査読無、2008、110-119

④ 佐々木 稔、堀口 正之、蔵野 正美、“区間ベイズ推定による適応型品質管理”、京都大学数理解析研究所講究録 1589「不確実な状況における意思決定の理論と応用」、査読無、2008、120-129

⑤ 堀口 正之、“マルコフ決定過程における適応型アルゴリズム (Adaptive Algorithms for Markov Decision Processes)”、査読無、神奈川大学工学研究所所報、2008、22--29

⑥ T. Iki, M. Horiguchi, M. Kurano, “A structured pattern matrix algorithm for multichain Markov decision processes”, Mathematical Methods of Operations Research, 査読有, Vol.66, 2007, 545-555

⑦ T. Iki, M. Horiguchi, M. Yasuda, M. Kurano, “A learning algorithm for communicating Markov decision processes with unknown transition matrices”, Bulletin of Information and Cybernetics, 査読有, Vol.39, 2007, 11-24

⑧ T. Iki, M. Horiguchi, M. Yasuda, M. Kurano, “Temporal Difference-Based Adaptive Policies in Neuro Dynamic Programming”, In: 4th International conference on Proceedings of Modeling Decisions for Artificial Intelligence (MDAI) 2007 (CD-ROM Proceedings), Vicenç Torra, Yasuo Narukawa, Yuji Yoshida (Eds.), 査読有, 2007, 112-122

⑨ 堀口正之、蔵野正美、安田正實, “マルコフ決定過程におけるTD法による学習アルゴリズムについて(A learning algorithm of TD method for Markov decision processes)”, 京都大学数理解析研究所講究録 1559「最適化問題における確率モデルの展開と応用」, 査読無, 2007, 34--49

〔学会発表〕(計６件)

① 堀口正之、“Uncertain Markov decision processes and Bayesian intervals”、日本数学会２０１０年度年会統計数学分科会、２０１０年3月２６日、慶應義塾大学

② 発表者：堀口正之、共同研究者：安田正實、“On bounds for Bayes estimate intervals in uncertain MDPs”、日本数学会２００９年度秋季総合分科会、２００９年9月２７日、大阪大学

③ 発表者：堀口正之、共同研究者：伊喜哲一郎、安田正實、蔵野正美、”Bayesian approach to uncertain MDPs with intervals of prior measures”、 日本数学会２００９年度年会統計数学分科会、２００９年3月２７日、東京大学

④ 発表者：堀口正之、共同研究者：伊喜哲一郎、蔵野正美、安田正實、”Adaptive algorithm for MDPs using pattern matrix learning method”、 日本数学会２００８年度秋季総合分科会統計数学分科会、２００８年9月２７日、東京工業大学

⑤ 発表者：堀口正之、共同研究者：伊喜哲一郎、蔵野正美、安田正實、”Adaptive Markov decision processes based on temporal difference method”、 日本数学会２００７年度秋季総合分科会統計数学分科会、２００７年9月２４日、東北大学

⑥ 発表者：堀口正之、共同研究者：伊喜哲一郎、”未知の推移法則を持つマルコフ決定過程における学習アルゴリズムについて”、日本数学会第１１７回九州支部例会、２００７年１０月１３日、宮崎大学

〔その他〕
ホームページ等
http://www.math.kanagawa-u.ac.jp/~horiguchi

６．研究組織
(1)研究代表者
堀口　正之 (HORIGUCHI MASAYUKI)
神奈川大学・工学部・准教授
研究者番号：９０３６６４０１

(2)研究分担者
(　　)

研究者番号：

(3)連携研究者
(　　)

研究者番号：