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 7 日現在

機関番号：８２６４６
研究種目：基盤研究（C）
研究期間：2010～2012
課題番号：２２５００１４３
研究課題名（和文）　経験強化型学習 XoL に関する発展的研究

研究課題名（英文）　Progressive research on the exploitation-oriented learning XoL

研究代表者
宮崎　和光（MIYAZAKI　KAZUTERU）
独立行政法人大学評価・学位授与機構・研究開発部・准教授
研究者番号：20282866

研究成果の概要（和文）： 得られた経験を強く強化する機械学習手法である「経験強化型学習 XoL」の発展として、「複数種類の報酬と罰を扱える手法」を完成させるとともに、応用の際に特に重要となる「報酬と罰の設計指針」の提示に成功した。具体的な応用例として、「科目の分類を支援する実システム」、「２足歩行ロボットの腰軌道学習」および「Keepaway タスクと呼ばれるサッカーを模したゲーム問題」への適用を行った。これらの成果により、伝統的な強化学習手法に対する XoL の優位性を強く主張できたと考える。

研究成果の概要（英文）：This research has completed an Exploitation-oriented Learning (XoL) method that can treat multiple rewards and penalties. Furthermore the design guideline of rewards and penalties on the XoL method has been proposed through illustrative examples, namely, a course classification task, a waist-trajectory learning task for a tendon-driven biped robot, and a Keepaway task in a multi-agent environment. It claim that XoL surpass traditional Reinforcement Learning based on Dynamic Programming in application to real-world problem.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	1,000,000	300,000	1,300,000
２０１１年度	1,000,000	300,000	1,300,000
２０１２年度	1,000,000	300,000	1,300,000
年度			
年度			
総　計	3,000,000	900,000	3,900,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：経験強化型学習、強化学習、報酬と罰の設計指針

１．研究開始当初の背景

今日、世界中のさまざまな領域において、機械の知能化による高付加価値化は必要不可欠であると言える。とりわけ、介護・子守ロボットなど、人と直接触れ合うことが要求される機械には、人の気持ちが理解できる高度な知能（優しさ）が必要とされる。そこでは、既知の部分は予めプログラミングすることができるが、それ以外の未知な部分は、人との触れ合い等を通して、試行錯誤的に、機械自らが学習し、与えられた目的（任務）を遂行することが求められる。

このような「試行錯誤に基づく目的指向の学習」は、現在、「強化学習」において集中的に研究されている。強化学習では、結果の善し悪しのみを表す「報酬」および「罰」と

いう弱い教師信号により学習が進行する。正解を与えることなしに学習できる点がたいへん魅力的な枠組みではあるが、現状では、キラーアプリケーションと呼べるような応用例が極めて少ないのも事実である。

これには主として次のふたつの点が関係していると考える。第一に、学習に時間がかかりすぎる点、第二に、適切な学習をさせるための報酬（罰）の設計指針が存在しない点である。これらはともに、理論的な解析においては問題とされない場合が多いが、手法の応用を考えた場合、深刻な問題を引き起こす。

これに対し、研究代表者らは、報酬（罰）に値を設定せずに、あくまで目的達成時（制約違反時）の信号として扱う立場をとっている。また、得られた経験を強く強化することで、試行錯誤回数の軽減をも目指している。現在、これらの特徴を満たす接近法として「Exploitation-oriented Learning (XoL)」を提唱している。

XoL は、これまでの強化学習とは異なり、手法の応用を主眼に置いた接近法であるが、応用の際に重要となる「報酬と罰の設計指針」は明らかにされていない。また、取り扱い可能な報酬と罰の種類にも制限がある。本格的な応用例の拡大を図るためにも、これらの点を早急に解決することが待ち望まれている。

２．研究の目的

本研究課題の目的は、

(1) 連続入出力に対応した XoL 手法の提案、
(2) 複数種類の報酬と罰に対応した XoL 手法の提案、
(3) XoL の応用例の提示、

および

(4) XoL における報酬と罰の設計指針の提示

である。

３．研究の方法

本研究課題は、研究代表者が 1994 年に証明した Profit Sharing と呼ばれる強化学習手法に関する定理に端を発する。その後、多くの研究を経て、平成 21 年には、XoL の中心となる手法である PS-r#を完成させ、XoL の枠組みを固めた。

PS-r#は、「離散的な状態-行動空間」および「１種類の報酬」を前提としている。実問題では、連続的な入出力への対応が要求される可能性が高い。そこで、まず初めに、PS-r#の「連続入出力への対応」を行う。

また、XoL を用いた応用例の拡大には、取り扱い可能な報酬および罰の種類を増大させる必要がある。研究代表者は、これまでにも、複数種類の罰を扱うプロトタイプ的な手法を提案しているが、本研究課題では、それを発展させ、「複数種類の報酬と罰に対応した XoL 手法」を完成させる。

その後、「連続入出力に対応した手法」と「複数種類の報酬と罰に対応した手法」とを組み合わせ、応用例の拡大を図る。提示した応用例を比較検討することで、「XoL における報酬と罰の設計指針」として取りまとめる。

４．研究成果

(1) 連続入出力に対応したXoL手法を提案した。具体的には、2007年に提案した連続入力に対応した罰回避政策形成アルゴリズムに対し、連続行動に適した行動選択方法を組み合わせることで、多様な行動の生成を可能にした。倒立振子の振り上げ安定化問題に適用することで、提案手法の有効性を確認した。

このことは、報酬と罰が各々高々１種類の場合に対応したXoLの基本的手法が完成したことを意味する。

(2) 複数種類の報酬と罰に対応したXoL手法の提案を行うとともに、連続値入力に対応したXoL手法との組み合わせに成功した。これにより、「連続値入力に対応した複数種類の報酬と罰を扱える手法」が完成した。

このことは、(4)で述べる「XoLにおける報酬と罰の設計指針」提示のための基本手法が完成したことを意味する。

(3) XoLの応用例として、以下の３項目を実施した。

① 独立行政法人 大学評価・学位授与機構における科目分類支援システムへの適用を行った。専攻の区分のひとつである「情報工学」区分を対象に、実際に申請されたデータを用いて、XoLの有効性を検証した。

本応用例は、現実の問題へのXoLの応用例として特に重要なものであり、今後実施する予定である「データベース作成・更新機能の実現」及び「様々な専攻の区分における有効性の検証」につなげるための準備を整えることができた。

② 腱駆動方式の二足歩行ロボットの腰軌道学習への適用に成功した。静的歩行を行うロボットの行動に、XoLが学習した出力を加えることで、直進動作の動的歩行を実現した。その際、「固定状態の導入」および「行動数の削減」が、学習の高速化に大きく寄与することを示した。実ロボットのシミュレータを作成し、有効性の検証を行った。

本応用例は、ロボット制御へのXoLの応用として非常に重要なものである。今後は、「実

ロボットを用いた検証」および「階段昇降等の様々な歩容の実現」へと研究を発展させる予定でいる。
③ 複数台のサッカーロボットによるパス回しを模したタスクであるKeepawayタスクでの検証に成功した。連続的な入力と離散的な入力が混在する場合の取り扱い方法について詳細に検討することで、代表的な強化学習手法であるSarsaに対するXoLの優位性が主張できた。シミュレーションによる検証を行った後に、LEGOロボットによるXoLの基本的な学習能力の確認を行った。

本応用例は、マルチエージェント環境下へのXoLの応用として重要な位置を占めるものである。今後は、「LEGOロボットを用いたKeepawayタスクの本格的な検証」を実施する予定でいる。

(4) (3)で述べた応用例を通じて、「XoLにおける報酬と罰の設計指針」の提示に成功した。具体的な設計指針を、計測自動制御学会の学会誌である「計測と制御」誌に解説記事として取りまとめ執筆した。これにより、試行錯誤に基づく学習手法としての XoL の存在意義を強く主張できたと考える。

以上の成果は、それぞれ、国内の学会や国際会議で発表し、高い評価を得た。また、国際的な学術雑誌へ掲載され、広く国際的にアピールすることにも成功している。

今後の課題としては、より広範囲な問題クラスへの適用と、さらなる手法の洗練化が挙げられる。

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 17 件）

① 宮崎和光, リレー解説 強化学習の最近の発展「第５回：応用志向の試行錯誤に基づく目的指向学習」Exploitation-oriented Learning; XoL, 計測と制御, 査読無, Vol.52, No.5, 2012, pp.462-467.
② Seiya Kuroda, Kazuteru Miyazaki and Hiroaki Kobayashi, Introduction of Fixed Mode States into Online Reinforcement Learning with Penalty and Reward and Its Application to Waist Trajectory Generation of Biped Robot, Journal of Advanced Computational Intelligence and Intelligent Informatics, 査読有, Vol.16, No.6, 2012, pp.758-768.
③ Kazuteru Miyazaki, Propocal of the Continuous-Valued Penalty Avoiding Rational Policy Making Algorithm, Journal of Advanced Computational Intelligence and Intelligent Informatics, 査読有, Vol.16, No.2, 2012, pp.183-190.

〔学会発表〕（計 11 件）

① Kazuteru Miyazaki, Proposal of an Exploitation-oriented Learning Method on Multiple Rewards and Penalties Environments, The 2nd International Conference on Applied and Theoretical Information Systems Research (2nd ATISR), 2012年12月29日, 圓山大販店, 台湾.
② 宮崎和光, 複数種類の報酬と罰に対応した経験強化型学習の提案と設計指針に関する研究, 平成 24 年度 電気学会 電子・情報・システム部門大会, 2012 年 9 月 7 日, 弘前大学.
③ Kazuteru Miyazak, Proposal and Evaluation of the Active Course Classification Support System with Exploitation-oriented Learning, The 9th European Workshop on Reinforcement Learning (EWRL-9), 2011年9月11日, Athens Royal Olympic Hotel, ギリシャ.

〔図書〕（計 1 件）

① Kazuteru Miyazaki, Exploitation-oriented Learning XoL - A new approach to machine learning based on trial-and-error searches – (Chapter 15), Multi-Agent Applications with Evolutionary Computational and Biologically Inspired Technologies : Intelligent Techniques for Ubiquity and Optimization, Kambayashi, Y. (Ed.), IGI Global, 2010, pp.267-293.

〔産業財産権〕
○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕
ホームページ等
http://svrrd2.niad.ac.jp/faculty/teru/xol_s.html

６．研究組織
(1)研究代表者

宮崎　和光（MIYAZAKI　KAZUTERU）
独立行政法人大学評価・学位授与機構・研究開発部・准教授
研究者番号：20282866