様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　２７　日現在

機関番号：　１２１０２
研究種目：　若手研究(B)
研究期間：　2012～2015
課題番号：　２４７６０３０８

研究課題名（和文）報酬が動的に変化する環境における事前知識を活用する強化学習

研究課題名（英文）Reinforcement Learning for Environment with Dynamic Reward using Prior Knowledge

研究代表者
澁谷　長史（Takeshi, Shibuya）

筑波大学・システム情報系・助教

研究者番号：９０５８２７７６

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：試行錯誤のさなかに報酬が変化する環境において、変化に関する事前知識を活用した効率的な強化学習方法を実現することである。強化学習には、多くの潜在的アプリケーションが期待されている反面、「ある行動を選択することの望ましいは時間に対して不変である」という仮定があり、時間とともに性質が変化する対象を学習できないという本質的な課題がある。本研究では、この実現の切り口として、変化に関する事前知識を活用した手法について研究を進めた。環境にあわせた事前知識を用いて、時間に対して報酬が周期性を持つ環境、方位に対して報酬が周期性を持つ環境、状態遷移確率が変化する環境などのための学習方式を明らかにした。

研究成果の概要（英文）：The purpose of this study is to develop efficient reinforcement learning method using prior knowledge for dynamic environment. Because conventional reinforcement learning method assumes that the environment is static, it is hard to be learned.
This study focus on prior knowledge to overcome this difficulty and proposes earning method for environment which has periodicity according time, direction and whose transition probabilities varies according time.

研究分野：　機械学習

キーワード：　強化学習

１．研究開始当初の背景

自ら行動し経験を重ねることで振る舞いを獲得する機械学習の枠組みとして強化学習がある。設計者は望ましい行動に対して正の報酬を、望ましくない行動に対して負の報酬をあらかじめ設定しておき、学習の主体であるエージェントに多くの報酬を集める行動を獲得させることで、設計の意図を間接的にエージェントに伝えることができる。

しかし、強化学習には、「エージェントが時間に対して変化する対象を学習できない」という本質的課題がある。これは強化学習が、時間に依存した変化を記述する能力がないマルコフ決定過程を基礎においているためである。マルコフ決定過程とは、エージェントの状態遷移や、エージェントに与えられる報酬が、現在の状態や行動にのみ依存して確率的に決定されるモデルであるが、現実問題として、実環境をマルコフ決定過程で記述することが困難である場合は少なくない。特に、報酬が変化する場合の研究については、あまり例が見られない。報酬が変化する場合の学習を可能にするということは、次のふたつの点で重要な意味を持つ。

まずひとつは、目標が変化する行動を学習できるようになることである。たとえばロボットの歩行制御などでは、マクロな視点では、初期位置から目標位置まで移動するという静的な問題としてとらえることができる。しかし、ミクロな視点では、一瞬一瞬の望ましいポーズは時間に依存して異なっているため、目標が変化する問題ととらえることができる。

もうひとつは、トータルの学習時間を短縮できることである。通常、報酬が変化した場合には、変化したあとの環境において学習を一からやりなおさなければならない。報酬は設計者が自身で設定するものである。

この点は、状態遷移が変化する場合の研究とは異なっている。すべてを学習するのではなく、使える知識は積極的に利用すべきである。

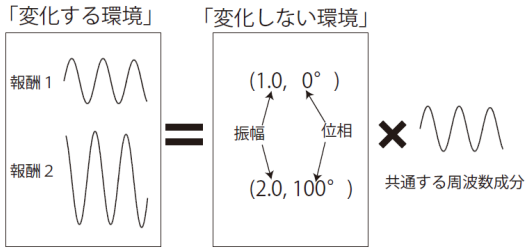

図１　報酬が変化する環境の分解

この観点にたち、これまで代表者は、報酬が周期的に変化する場合と非周期に変化する場合のふたつの場合について、検討を行ってきた。周期的な変化については、電気回路の分野で変動を扱う手法として知られている「フェーザ表示」を用いる手法を提案してきた。この手法は、報酬を既知の周波数をもつ正弦波の電圧源に見たて、学習対象の環境を、時間変動成分を取り除いた振幅と位相のみからなる環境に変換するというものである（図１）。非周期な変化についても、報酬を線形関数の和として記述できる場合について検討をおこなってきた。

これらの検討は、ふたつの目標が入れ替わるという限定的な環境においてその有効性が確認されているのみであり、より大きな、より複雑な環境において検証し、手法が有効に働く問題の領域を実験的に特定する必要がある。

２．研究の目的

本研究の目的は、試行錯誤のさなかに報酬が変化する環境において、変化に関する事前知識を活用した効率的な学習方法を実現するための基礎理論を整備することである。

３．研究の方法

この目的のために、代表者がこれまで実施してきた報酬が変化する環境のための学習方法を主な基礎として、次の方法で、研究を進めた。

(1) 時間に対する周期性をもつ環境での学習法の開発

(2) 方位に対する周期性環境での学習法の開発
(3) 位置に対する周期性環境での学習法の開発
(4) 故障などの環境変化に対応する学習法の開発
(5) 連続空間における学習法の開発

４．研究成果

本研究における具体的な成果は以下の通りである。

(1) 時間に対する周期性をもつ環境での学習法の開発

報酬が周期的に変化する環境のための強化学習法を提案した。変化の周波数に関する事前知識を活用することによって、学習すべき行動価値関数を、複数の既知の正弦波数と、周波数ごとの未知の係数から構成できることが明らかになり、これによりこのような環境における学習の効率化を図ることができる。提案手法では、時間の周期関数としての報酬をフーリエ級数として分解し、行動価値関数の各周波数における振幅と位相を学習する。数値実験の結果、提案手法は、従来手法よりも学習にかかる試行数を減少させる効果を確認した。

(2) 方位に対する周期性環境での学習法の開発

(1) の成果を土台として、方位に対する周期性に着目して、学習にかかる試行数を低減させる学習法について提案した。数値実験の結果、提案手法は、従来手法よりも学習にかかる試行数を減少させる効果を確認した。

(3) 位置に対する周期性環境での学習法の開発

状態空間内の周期運動を学習する方式について検討し、(1)の方法を発展させることで、これを実現できることが示唆された。

(4) 故障などの環境変化に対応する学習法の開発

故障を起こさないために安全度を予測しながら学習する方法や、故障を起こしたあとにその身体での高い性能を早く発揮するための学習方法、また、状態遷移確率の変化を事前知識として用いることができる学習方法を提案し、それぞれ数値実験で有効性を確認した。

(5) 連続空間における学習法の開発

連続空間における学習方法について、関連する研究者と議論を行った。選択的不感化ニューラルネットワークにより行動価値関数を近似する手法を提案し、学習性能が向上することを確認した。

５．主な発表論文等

〔雑誌論文〕（計 6 件）

(1) Moriaki Onishi and Takeshi Shibuya, A study of efficient reinforcement learning using the relative angle of two objects, Proceedings on the 16th International Symposium on Advanced Intelligent Systems, pp.1091-1098, 2015(査読あり)
http://www.isis2015.org/main/

(2) Junki Tamaru and Takeshi Shibuya, Profit Sharing reducing the occurrences of accidents by predicted action-safety degree, Proceedings of the 10th Asian Control Conference 2015 (ASCC 2015), pp.2468-2473, 2015(査読あり)
http://dx.doi.org/10.1109/ASCC.2015.7244700

(3) 小林高彰，澁谷長史，森田昌彦，選択的不感化ニュ-ラルネットを用いた連続状態行動空間におけるQ学習，電子情報通信学会論文誌 D，Vol.J98-D，No.2，

pp.287-299, 2015(査読あり)
http://search.ieice.org/bin/summary.php?id=j98-d_2_287

(4) Takeshi Shibuya, Reinforcement learning using BAMDP-based prior knowledge for dynamic environment, USB Proceedings of the 11th International Conference on Modeling Decisions for Artificial Intelligence, pp.143-152, 2014(査読あり)
http://www.mdai.cat/mdai2014/

(5) Takaaki Kobayashi, Takeshi Shibuya and Masahiko Morita, Q-learning in Continuous State-Action Space with Redundant Dimensions Using a Selective Desensitization Neural Network, Proceedings of Joint 7th International Conference on Soft Computing and Intelligent Systems and 15th International Symposium on Advanced Intelligent Systems, pp.801-806, 2014(査読あり)
http://dx.doi.org/10.1109/SCIS-ISIS.2014.7044714

(6) 澁谷長史, 安信誠二, 報酬が周期的に変化する環境のための強化学習, 電気学会論文誌C, Vol.134, No.9, pp.1325-1332, 2014(査読あり)
http://doi.org/10.1541/ieejeiss.134.1325

〔学会発表〕（計　11　件）

(1) NGUYEN VAN BAC, 澁谷長史,外部による評価を報酬に組み入れる繰り返し動作の獲得手法の一検討, 電気学会システム研究会, 2015/12/06, 新潟県立看護大学(新潟県上越市)

(2) 羽鳥貴久, 澁谷長史,フレーム変形したロボットのための事前学習による効率的な動作獲得法の検討, 第42回知能システムシンポジウム, 2015/03/18, 北野プラザ六甲荘(兵庫県神戸市)

(3) 臼井翼, 澁谷長史,事前知識を反映した状態遷移確率推定により環境変化に適応する強化学習, 第41回知能システムシンポジウム, 2014/03/13, 筑波大学東京キャンパス(東京都文京区)

(4) 澁谷長史, 報酬を与えられる領域が変化する環境における強化学習, 平成２４年 度電気学会電子・情報・システム部門大会, 2012/09/07, 弘前大学文京町キャンパス(青森県弘前市)

(5) Takeshi Shibuya, Profit Sharing reducing the occurrences of accidents by predicted action-safety degree, 10th ASIAN Control Conference, 2015/06/03, コタキナバル（マレーシア)

〔その他〕
ホームページ等
http://www.mil.iit.tsukuba.ac.jp/

６．研究組織
(1)研究代表者
澁谷 長史　（SHIBUYA, Takeshi）
筑波大学・システム情報系・助教
研究者番号：　９０５８２７７６