様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　　元 年　　６ 月 ２７ 日現在

機関番号：３３９１９
研究種目：若手研究(B)
研究期間：2016～2018
課題番号：１６Ｋ１６１３２

研究課題名（和文）非同一マルコフ決定過程間での徒弟学習によるロボットの行動学習

研究課題名（英文）Apprenticeship learning for heterogeneous robots

研究代表者
増山　岳人（Masuyama, Gakuto）

名城大学・理工学部・准教授

研究者番号：２０７０７０８８

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：非同一な身体性，環境をもつ二者間で報酬関数を転移する逆強化学習手法と，関連する基礎技術について研究を実施した．特に有用性が期待できる研究成果は以下の2つである．1)身体性，環境が異なることで両者から観測される軌道のなす特徴量の時系列に不整合性が生じるという問題に対し，事前に与えられた対応点を利用して特徴写像を陰に学習し，エキスパートから与えられる演示軌道を学習者の特徴空間上で近似する手法を開発した．2)演示に限らず任意の軌道に対するスコアから非線形な報酬関数を推定するアルゴリズムを提案した．

研究成果の学術的意義や社会的意義
人手で目的関数を設計することなく，観測情報に基づいてロボット単体で目的関数を構成することは，ロボットの自律性向上という意味で意義があるものと考える．現在の技術で目的関数を推定するには，何らかのお手本となるデータをロボットに観測させる必要があるが，一方で観測する対象とロボットでは身体，社会から求められる要請など，多くの差異がある．そのため，単純な模倣の枠組みでは適用可能な場面が限られる．本研究課題ではこの問題を緩和する新たな知見を提示した．

研究成果の概要（英文）：What this project pursued is to develop algorithms that transfer reward function between heterogeneous agents. Relevant inverse reinforcement learning techniques were also studied. Representative contributions of this project are as follows: 1) Inverse reinforcement algorithm assuming that an expert and agent follows non-identical Markov decision processes, or incompatible features. To represent demonstrations of expert observed in distinct feature space, a conditional density estimation technique is leveraged, and it is shown that approximation of demonstrations in agent feature can be represented in closed-form with a specific model. 2) Non-linear score-based inverse reinforcement learning, which enables us to use arbitrary trajectories, i.e. trajectories sampled from pre-learned policy of an agent, to estimate reward function.

研究分野：知能ロボティクス

キーワード：徒弟学習　逆強化学習

様　式　C－１９、F－１９－１、Z－１９、CK－１９（共通）

１．研究開始当初の背景

強化学習は，将来に渡って観測される報酬信号の期待値を最大化する方策を学習する，汎用的な最適化の枠組みである．ロボットに与えられるタスクを表現する報酬関数は，方策学習のサンプル効率に影響するため，その設計は慎重になされなければならない．しかし，報酬関数の設計は非常に煩雑な作業であることが多く，設計基準が曖昧にならざるを得ないタスクも少なからず存在する．

報酬関数の設計問題に対しては，所与のタスクに対するエキスパートの演示から，その方策を与える報酬関数を推定する徒弟学習（逆強化学習）が考案されている．徒弟学習はエキスパートの演示に基いてその制御方策を学習するため，模倣学習の一種であるといえる．報酬関数の推定と順強化学習によって制御方策を学習するため，演示が観測されない領域においても報酬期待値を最大化する制御方策が得られるという特長がある．

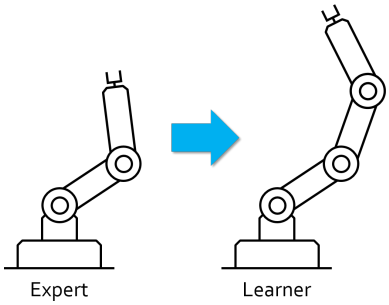

図　1　非同一ロボット間での報酬関数転移

徒弟学習では，一般的にエキスパートとロボットが同一の問題設定（Markov Decision Process; MDP）を扱うことが前提とされる．エキスパートとロボットの振る舞いは単一の MDP によって記述されるが，一般に両者の間には 1)身体の差異，及び 2)環境の差異が存在する（図 1）．身体及び環境の差異は不可避に発生するが，従来の徒弟学習では事前に設計者が十分な注意を払い，問題毎に適切な前処理を行うことでこれに対処している．しかし，推定された報酬関数に基づき，現在の MDP でのサンプルから制御方策を汎化する徒弟学習の枠組みには，オンラインでの優れた適応能力が期待できる．そのため，従来の徒弟学習の枠組みは，その特長を十分に活用しきれていないと考える．

２．研究の目的

身体及び環境がロボットとは非同一であるエキスパートの演示から，適切な報酬関数を推定する徒弟学習手法の開発を目的とし，
（1）エキスパートとロボット間での特徴写像の学習，
（2）サンプル軌道の自己生産による報酬関数推定，
（3）軌道間距離尺度の設計，
以上，3 つの研究課題を設定した．

３．研究の方法

（1）エキスパートとロボット間での特徴写像の学習

一般に徒弟学習では，エキスパート及びロボットそれぞれの方策から観測される特徴量の期待値が合致するよう報酬関数の推定を行う．したがって，エキスパートの特徴期待値をロボットの特徴空間へ写す関数が得られれば，逆強化学習の適用自体は可能になる．

しかし，エキスパートとロボットの間で同一の特徴を用いることが現実的ではない場合，特徴期待値のマッチングを行うことはできない．そこで，事前に用意された対応点のデータセットから 2 つの特徴空間の間の写像を学習することでこの問題を解決し，徒弟学習の枠組みの適用範囲の拡大を図った．

（2）サンプル軌道の自己生産による報酬関数推定

上述の特徴写像を学習することで報酬関数の推定は可能となるが，エキスパートの演示に対応する軌道がロボットによって常に実行可能であるとは限らない．

そこで，エキスパートの演示を直接報酬関数推定に用いるのではなく，実行可能であることが保証されるロボット自身によってサンプルされた軌道のうち，演示との類似度が高いものを新たな演示として利用する方法論を提案した．

（3）軌道間距離尺度の設計

自己生産したサンプル軌道から新たな演示を選ぶとき，エキスパートの演示とサンプル軌道の類似度をどのような尺度で選べばよいかという問題は自明ではない．MDP 間に不整合性が存在するとき，ナイーブな特徴量の空間的距離尺度では有用なサンプル軌道を適切に評価できない場合がある．

そこで，軌道をなす特徴量の時系列をさらに抽象化した特徴量を導入する．具体的には，特徴系列に対するアフィン変換に関する不変量を用いる．

（4）能動学習による対応点の選択

（1）で述べた特徴写像の学習では，エキスパート及びロボットの特徴の対応点を複数必要と

する．しかし，対応点の収集には多大な労力を要することが想定され，本研究の応用上でのボトルネックとなり得る．

そこで，必要な対応点数を最小限に抑えるため能動学習を用いる．特に精度よく写像を学習する必要があるのは，報酬関数推定に関連するエキスパートの演示及び対応するロボット特徴空間における軌道周りである．そこで，はじめに少ない対応点で大雑把な写像を学習し，その後に写像された特徴空間における演示軌道周りについて対応点のクエリを出す．

４．研究成果

（1）特徴写像の学習と自己生産した軌道を用いた報酬関数の推定

異なる特徴空間の間の特徴写像を所与の対応点から学習することで報酬関数を推定する逆強化学習手法を開発した，学習者と異なる MDP にしたがうエキスパートの演示を，学習者の特徴空間において表現することができればよいが，これは一般にサンプル効率，精度の面で容易ではない．

そこで，学習者の特徴空間における演示の確率密度を周辺化するというアプローチをとった．確率密度の推定もまた困難であるため，実際に推定するのは確率密度比とし，学習者の特徴空間における演示をエキスパートの演示から直接近似する．また，近似モデルをガウシアンとする場合には解析的に表現可能であることを示した．シミュレーション結果より，エキスパートと学習者で身体性，環境が異なる場合にも適切に報酬関数が転移可能であることを確認した．図 2 に提案手法の概要を示す．

一方，興味深い結果として，周期的な運動を含むような問題では，エキスパートとロボットの身体性が大きく異なっていたとしても徒弟学習が適用可能である場合が確認できた．これは人とロボットというよりも，別種のロボット間での知識転移の可能性を示唆するものとも考えられる．この点について，追加調査を行っていく予定である．

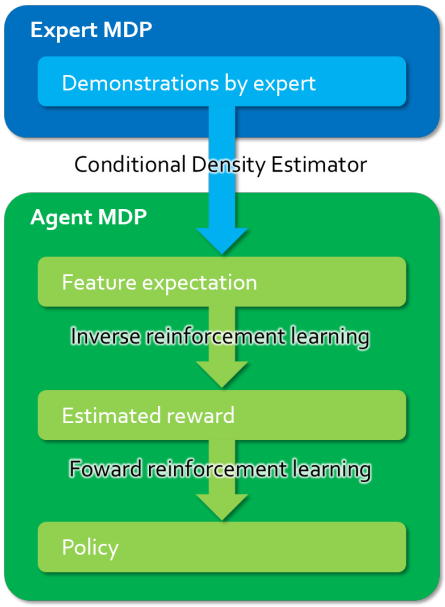

図 2 特徴写像を介した報酬関数推定手法の概要

（2）軌道間の類似度に関する研究

自己生産した軌道と演示の距離尺度としてアファイン変換不変量を導入した．各軌道がなす特徴空間における時系列から不変量を算出し，自己生産したサンプル軌道と演示の類似度を設計した．図 3 に，アファイン変換不変量を用いた軌道間類似度による報酬関数推定の手順を示す．

これにより，単純に演示軌道を用いた場合には徒弟学習の効果がほとんど得られないような問題に対しても適切な報酬関数を推定可能であることを確認できた．

しかし，どのような提案手法の適用が有効であるタスクを判別することが難しいという問題がある．また，用いる特徴量によっても有効性が変わるが，その設計指標も明らかでないという課題が残っている．本研究課題の実施期間中に．上記の問題を解決することは難しいと判断し，アプローチを再検討することとなった．

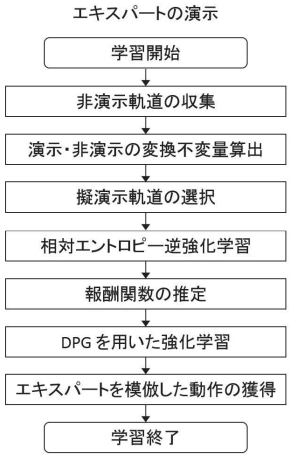

図 3 軌道間類似度に基づく逆強化学習

（3）スコアに基づく報酬関数推定

徒弟学習，逆強化学習の応用上の問題点の１つは，報酬関数の推定に多くの演示データが必要となることである．これは実データを扱う際のコストを増大させる要因であり，報酬関数設計のプロセスを簡単化したいという従来のモチベーションと反する結果にもなり得る．そこで，エキスパート側と学習者側とのインタフェースとして演示以外のチャネルを用意することが重要であると考えた．

そこで，軌道のスコアに対する逆強化学習に着目し研究を進めた．この手法では，最適な演示軌道の代わりに任意の軌道の ”よさ ” を表すスコアを用いて報酬関数を推定する．スコアは多くの場合，人が付与することが想定される．本研究では，能動学習の導入によるスコアあり

軌道の本数の低減及び報酬関数モデルの非線形化に取り組んだ．

まず，能動学習の利用に関しては，クエリ軌道の生成に必要な特徴量の累積値を計算するための動的計画法を提案した．ただし，方策ごとに動的計画法を実行することは困難であるため，行動は一様分布にしたがって決まるものとしている．図 4 は提案した動的計画法を用いたクエリ軌道生成器と軌道のスコアに基づく報酬関数推定器との関係を表している．

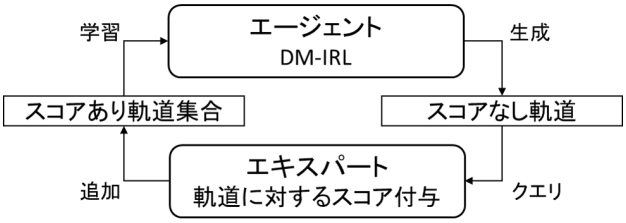

図 4 クエリ軌道の生成と報酬関数推定

クエリ軌道の生成基準には，推定スコアが最大となる軌道，スコアの推定誤差を利用した軌道，そして両者の組み合わせについて検証を行い，一定の効果があることを確認した．

報酬関数モデルの非線形化に関しては，任意の代表値によって与えられるカーネル関数を用いた定式化を行った．逆強化学習の表現能力を向上するとともに，パラメータ推定は任意の線形回帰により行うことが可能という特長をもつ．図 5 及び図 6 は，それぞれ線形及び提案した非線形な報酬関数モデルを用いて推定された報酬関数の一例である．

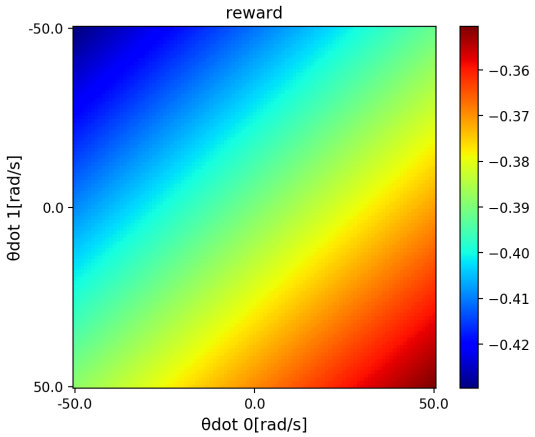

図 5 線形モデル

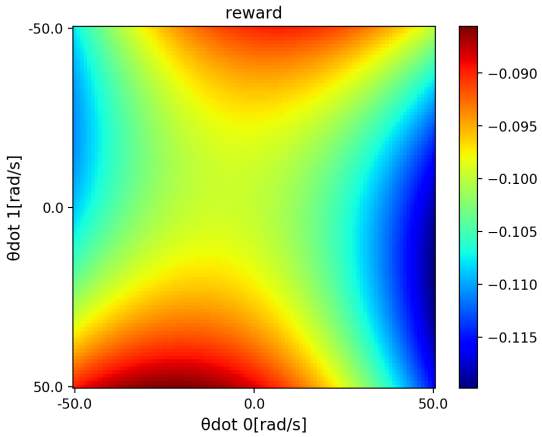

図 6 非線形モデル

ロボットの制御タスクを用いたシミュレーションを行った結果，線形モデルでは学習が困難なタスクでも，真の報酬関数を用いた場合と同等のパフォーマンスをもつ方策が学習可能であることを確認した．

また，強化学習のサンプル効率向上を目的に，方策学習の過程で生成される軌道を用いた報酬関数と方策を同時に学習する枠組みを構築した．現在の方策にしたがってサンプルされる軌道の一部に対するスコアをエキスパートに対するクエリとし，報酬関数と方策を交互に更新する．また，報酬関数を推定する回帰に探索を効率化するためのペナルティ項を加えた．ペナルティ項は，推定報酬関数に基づいて出力される軌道の推定スコアの分散から算出される．距離の近い軌道に対する評価の違いを強調することで探索を促進するものである．図 7 に提案手法

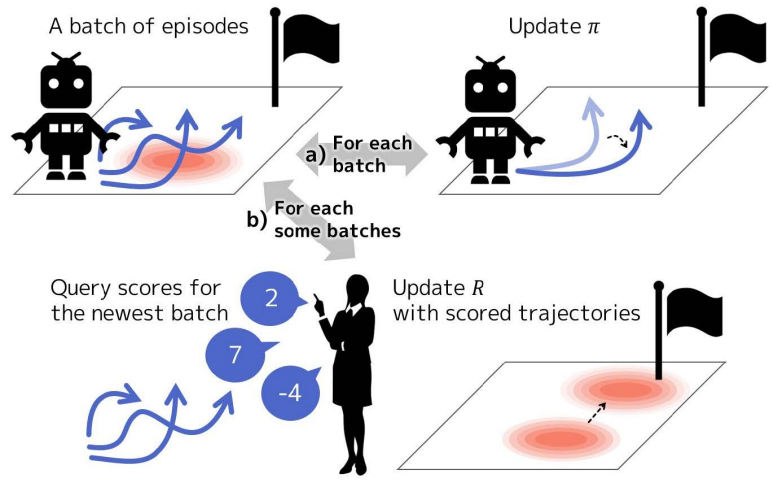

図 7 方策学習に用いた軌道に対するスコアのクエリ

の概念図を示す.
　シミュレーションの結果，特に方策学習の初期フェーズにおいて有効である可能性が示唆された.

５．主な発表論文等

〔雑誌論文〕（計　0　件）

〔学会発表〕（計　5　件）

[1] 徳永将典，増山岳人，多視点画像を用いたメタ学習による One-Shot 学習，日本機械学会ロボティクス・メカトロニクス講演会 2019, 2P2-I05, 2019.
[2] 久保匠，増山岳人，報酬値の信頼度とパラメータ分布間の KL ダイバージェンスを用いた進化戦略，日本機械学会ロボティクス・メカトロニクス講演会 2019, 2A2-B16, 2019.
[3] 渡邉夏美，増山岳人，梅田和昇，軌道のスコアに基づく逆強化学習を用いた非線形な報酬関数の推定, 2018 年度人工知能学会全国大会, 1N3-02, 2018.
[4] G. Masuyama, K. Umeda, Apprenticeship learning in an incompatible feature space, ICRA2017, pp.932-938, 2017.
[5] 渡邉夏美，増山岳人，梅田和昇，スコアに基づく逆強化学習のための動的計画法による軌道の自己生成，日本機械学会ロボティクス・メカトロニクス講演会 2017, 2P2-E05, 2017.

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年：
国内外の別：

○取得状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕
ホームページ等

６．研究組織

(1)研究分担者
研究分担者氏名：
ローマ字氏名：
所属研究機関名：
部局名：
職名：
研究者番号（8 桁）：

(2)研究協力者
研究協力者氏名：

ローマ字氏名：

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。