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

平成　３０　年　　６　月　２３　日現在

機関番号：　１２７０１
研究種目：　挑戦的萌芽研究
研究期間：　2016～2017
課題番号：　１６Ｋ１２４８５

研究課題名（和文）熟練者のスキルを超越するアンサンブル逆強化学習の提案

研究課題名（英文）An ensemble inverse reinforcement learning for exceeding the expert skills

研究代表者

濱上　知樹（HAMAGAMI, Tomoki）

横浜国立大学・大学院工学研究院・教授

研究者番号：３０３３４２０４

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：アンサンブル学習の一種であるAdaboostアルゴリズムを逆強化学習に応用し，複数の報酬関数を統合するシステムを構築した。目的のタスクを達成可能な方策を持つエージェント(サブエキスパート)を複数利用することで，よりタスクの学習に適した報酬関数を獲得するアンサンブル逆強化学習を開発した。複数のサブエキスパートから推定した報酬関数の統合により，それぞれの報酬関数に含まれる不完全知覚の影響緩和を狙いとして適切な報酬配分へのが可能になった。不完全知覚状態を含む環境における実験を行い，アンサンブル逆強化学習によってよりタスクの学習に適した報酬関数を獲得できること確認し，本提案システムの有効性を示した。

研究成果の概要（英文）：Ensemble inverse reinforcement learning from semi-experts' behavior is proposed. In many inverse reinforcement learning (IRL) problems, the expert agent which has ideal rewards for achieving the goal is supposed to be existing. However, in real-world problem, the expert is not always observed. Moreover, the estimated reward function includes the bias depending on its inherent behavior if the reward for achieving the goal task is estimated from one agent. In order to overcome the limitation of IRL, we apply Adaboost, one of ensemble and boosting approach, to IRL and integrate estimated reward functions from semi-expert agents. To confirm the effectiveness of the proposed method in the grid world including incomplete areas, we compared the results of reinforcement learning using estimated reward functions and integrated reward function by simulation. The simulation result shows the proposed method can estimate the reward adaptively.

研究分野：　情報学

キーワード：　逆強化学習　強化学習　アンサンブル学習　不完全知覚

## １．研究開始当初の背景

強化学習(Reinforcement Learning: RL)とは，報酬獲得を最大化する振舞いを，試行錯誤を通して獲得する能動学習である。環境適応が必要なシステムの設計に有効な学習手法として，これまで多くの研究事例がある。特に近年は深層化技術を組み合わせた深層強化学習が驚くべき成果を上げており，人工知能分野の中でも最も注目されている技術となっている。

しかし，現実問題の中には，学習者が獲得すべき報酬が陽に定義できない場合も多い。この問題に対しては，逆強化学習(Inverse RL: IRL)[1,2,3]が有効な解決策となる。IRL は，エキスパートと呼ばれる熟練者の振舞いを学習者が観察し，エキスパートが持つ報酬関数を推定しながら学習を進める。 Ng らはこの報酬推定問題を線形計画問題に帰着させ，複雑なヘリの操縦の学習に成功した。さらに，近年はベイズ推定に基づく報酬確率モデルや，連続系，MAS への拡張など，実問題への有効性が示されている。

一方，IRL にはエキスパートが所与であるという強い仮定が求められる。しかし，現実の問題ではエキスパートが陽に定義されない場合が多い。また，恣意的に定めたエキスパートからの IRL は，大きな学習バリアンス（分散）を引き起こすおそれがあり，汎化性能に限界がある。すなわち，IRL の実用上の課題は，容易に観測可能な多様な非エキスパートの振る舞いから，エキスパート並みかそれを超越する性能を引き出す報酬関数を推定しなくてはならない点にある。

## ２．研究の目的

本研究課題では，背景で述べた課題に対するブレイクスルーとして，複数の弱学習器を組み合わせる「アンサンブル学習」に着想を得た，独創的な逆強化学習の拡張法「アンサンブル逆強化学習（Ensemble IRL: e-IRL）」の理論展開と実現に取り組むを提案することを目的とする。具体的には，下記のことを明らかにすることをめざした。

(1) e-IRL の理論確立： e-IRL を実現する理論を確立し，その性能限界を明らかにする。

(2) e-IRL の実装評価： e-IRL を実装し，計算オーダーとスケーラビリティを評価する。

(3) e-IRL の応用実験： e-IRL を高度な自動運転タスクに応用し，有効性と性能を実証する。

逆強化学習をアンサンブルによって性能改善しようとする試みは過去に例のない独創的な試みである。また，エキスパートに頼らない新たな逆強化学習の提案は，他の手法にはみられない大きな特色である。さらに IRL のバイアス―バリアンストレードオフをはかる評価方法を明らかにすることで，推定された報酬関数の頑健性について学習前の評価が可能になる。すなわち，従来の IRL の性能と応用範囲を大きく拡げる波及効果が得られる。

e-IRL の活用範囲は極めて広い。特に期待できるのは自動運転の高度化である。多様な人(非エキスパート)による運転から推定される報酬関数を最適に組み合わせることで，協調的・適応的な高いスキルを獲得する。このように，知能システムの高度化に至る意義ある理論と結果を得る。

## ３．研究の方法

e-IRL の理論検討とこれを実装するためのアルゴリズムの開発に着手する。

※ アンサンブル逆強化学習の理論深化

e-IRL では，複数の非エキスパートから得られる報酬関数を適応的に組み合わせることで，環境変化に頑健な報酬関数の合成とこれを用いた効率的な強化学習を行う。このとき，アンサンブル学習からも推測される通り，非エキスパートのサンプリングがその性能に大きく影響を与えることが予想される。たとえば，問題空間の大きさに対するエキスパートのスキルを軸とする高次元空間を想定すると，非エキスパートのそれぞれはそのスキル空間中の部分空間に存在するであろう。その部分空間の集合が疎であっても密であっても，単なる寄せ集めでは最適な報酬関数は合成できないことは容易に想定できる。また，それぞれの非エキスパートが競合関係にあるマルチエージェント系の逆強化学習では，報酬関数同士が干渉しあうことで，最適な報酬関数の組み合わせは困難であることも明らかである。

以上の考察から，理論的検討の中心課題は，相互に干渉のある非エキスパートの最適な組み合わせという問題にあり，ここにアンサンブル学習である boosting を報酬関数の合成に活用する意義がある。この議論をもとに，報酬関数の boosting 理論をアンサンブル学習の方法を応用し発展させる。

次に理論検討された方法を実装する手段となるアルゴリズムを開発する。検討の基礎となるのは，研究代表者らがこれまで検討を進めてきた，スキルベース学習 e-IRL の基礎的検討である。

スキルベース学習は，非エキスパート同士のスキルが相互干渉する場合に，干渉による報酬の相殺を避けるためのサブ報酬関数を段階的に求めるアルゴリズムである。本研究においては，このアルゴリズムを基礎に，理論検討された最適化の方法に合わせて既提案方法を拡張することから開始し，実問題に応用可能な最適化を試みる。

また e-IRL の予備検討では，高度な振る舞い獲得に必要な報酬関数を，非エキスパートの報酬関数を適切に組み合わせて合成している。本研究においては，この基礎的アルゴリズムの最適化とスケール手法を模索することで大規模複雑な問題への適用を可能にする。

さらに，システムの上限性能や平均性能の向上だけではなく，下限性能を担保した学習方法をめざす。具体的には，問題のスケールと非エキスパートの有するスキル次元に対する計算コストの下限を担保するアルゴリズムの開発に取り組む。

## ４．研究成果

本研究の最も大き貢献果は，逆強化学習のアンサンブル化とそれによる熟練エキスパートの性能を超える性能の獲得である。そのために，Adaboost[4,5]アルゴリズムにヒントを得た逆強化学習の拡張を行った。この仕組みの概要を図１に示す。Adaboost では学習データに対する正誤判定によって学習器の信頼度を評価する。これに対応する処理として，本研究では特徴期待値の分布から外れ値検出を行い，各 sEA の振る舞いの信頼度を評価する。ここで信頼度は sEA の振る舞いを表す特徴期待値ベクトルを用いる。ここで，タスクの達成に不可欠な振る舞いが複数の sEA に共通して観測されるのであれば，特徴期待値の分散は小さくなる。そこで，特徴期待値の分布から外れ値検出を行い，これを用いて信頼度を評価する。

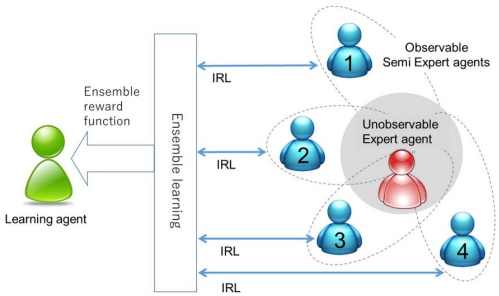


図１　アンサンブル逆強化学習の概要

以下に基本的なアルゴリズムを示す。

(1) K 体の sEA から Ng らによる見習い学習を行う。見習い学習は，強化学習を繰り返し行うことで徐々に EA の振る舞いに近づくよう Projection method を用いて報酬関数を修正する。図 2 に Projection method の模式図を示す。

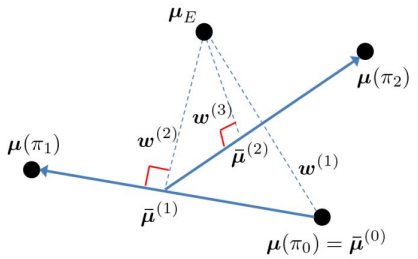


図２　Projection method

そして推定される報酬関数には最終的な目標状態や対応するゴール，またそこに至るまでのサブゴールに相当する報酬が設定される。

(2) それぞれの sEA の特徴に対し，分散と偏差を計算する。このとき，偏差に対して正規化を行う。

(3) 各特徴の重要度と信頼度を計算する。

(4) 重要度を外れ値判定を用いて更新する。

(5) すべての sEA に対し(1)~(4)の操作を行い，最後に信頼度の正規化後に報酬関数を統合する

以上の手続きを図３に示す。

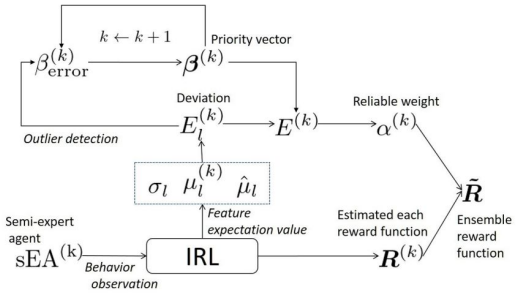


図３　弱エキスパートのアンサンブル化

これらの手順に従い，グリッドワールドにおける実験を行い，手法の有効性について評価をした。図４に実験に用いた実験環境を示す。

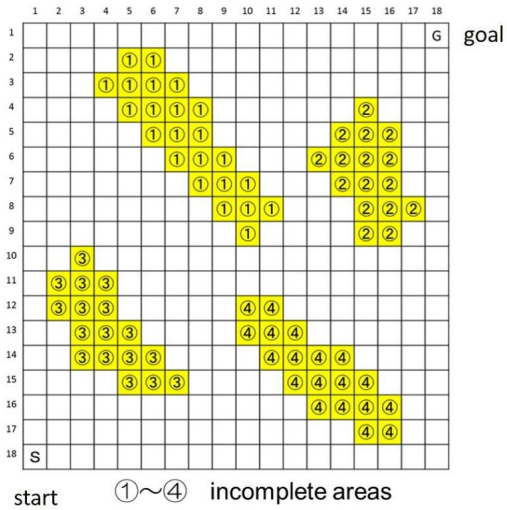


図４　不完全知覚のあるグリッドワールド

この実験では熟練エキスパートの能力を弱エキスパートのアンサンブルが超えることを示すため，S-G の経路生成が困難な不完全知覚問題を含む迷路を用いている。不完全知覚とは異なる状態を同じ状態のように認識してしまう問題である。図４中で同じ番号のセルは座標が異なっていても同じ状態のようにみえる。そのため，MDP を仮定した強化学習にとって，不完全知覚のセル内で最適行動をとることは難しい。これに対し，sEA はそれぞれ異なる不完全知覚の影響を受けているものとする。

一定性能の sEA を得るために，組み合わせ最適化問題ととらえ，遺伝的アルゴリズムによる 10 体の sEA を作成した。代表的な sEA の行動と価値関数を図５に示す。このように，部分的に不完全知覚の領域に入ってしまい，最適な行動が実現できていないことがわかる。他の９体もそれぞれ異なる不完全知覚の

影響を受けており，最適な行動獲得には至っていない。

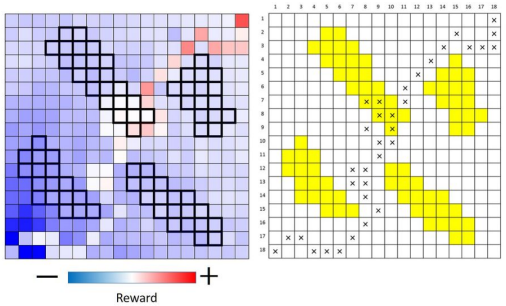


図５　sEA の典型的な行動および報酬関数

次に，各 sEA の行動をそれぞれ 100 エピソード分観測し，それぞれの特徴期待値を獲得した。強化学習の繰り返し回数を 50 回とし，予備実験によって学習率を 0.3，割引率を 0.99 に設定した。回あたりの学習ステップ数は 100000 回とし，価値更新アルゴリズムには Q-learning を，方策には $\epsilon$-greedy 選択を用いた。またランダム行動率 $\epsilon$ は 1.0 からスタートし，10000 ステップごとに 0.1 減少させ，$\epsilon = 0$ となってから，特徴期待値を 100 エピソード分観測した。１エピソードあたりの最大ステップ数は 1000 である。

図６にアンサンブル後のエージェントの動きと価値関数の分布を示す。図に示す通り，最適行動を実現する報酬関数推定に成功をしていることがわかる。

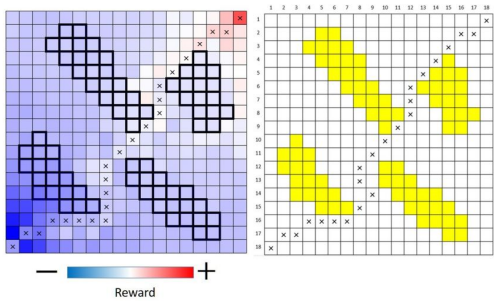


図６　アンサンブル逆強化学習結果

sEA とアンサンブルの結果を比較すると，いずれもゴールに近い状態ほど高い報酬が推定されていることがわかる。しかし，各 sEA から推定された報酬関数は，経路の途中に局所的な報酬の極大点が存在しているのに対し，アンサンブルされた結果はスタートからゴールまで経路に沿って単調に価値が増加している。さらに，アンサンブルされた結果は，障害エリアとこれに隣接したセルの価値が経路上のセルより低く推定されており，障害エリアの影響を受けにくい振る舞いが獲得できている。アンサンブルにおける sEA の提示順序は，得られる報酬関数および方策の性能に大きな影響を与えないことも明らかとなった。しかし，sEA の集合分布が異なる場合については，他のアンサンブル学習と同様に，提示順序がアンサンブル後の性能に影響することが予想される。

本研究に関連した IRL による報酬関数獲得法に関し，いくつかの条件を想定した類似研究が行われている。これらの研究との比較も行った。

荒井[2,3]らはマルチエージェント環境において，最適な行動系列を所与とした IRL を行い，最適なタスク配分を自律的に獲得する方法を提案している。これに対し本研究は，独立した相互作用のない複数の sEA に対し，それぞれから推測された報酬関数をアンサンブルすることで，平均的かつ多く観測される振る舞いを信頼しようとするものである。

桜井[7]らは，学習途上のエージェントからの複数の IRL を用いた報酬関数の推定法を提案している。本研究も複数の IRL の結果を用いた方策推定法であるが，一方で単独エージェントの学習に伴う変化への追従をめざした既存研究に対し，本研究は，それぞれの逆強化学習は独立した不完全な演示を行う複数の sEA を対象としている点が異なっている。

Choi[8]らは，本研究と同様に，POMDP 環境下における RL の理論を用いた IRL 拡張を図っている。本論文で扱ったグリッドワールドにおける障害エリアも部分的な不完全知覚状態を含んでいるが，POMDP 環境における直接的な解決を目的とはしていない。今回の研究においては，熟練者の能力を POMDP 環境での学習と置いて，アンサンブル逆強化学習を行ったが POMDP 環境における学習性能については，さらなる検討が必要である。特に，sEA の報酬関数の分布については一定の仮定が必要である。

Syed[9]らは，見習い学習において，ゲーム理論に基づく報酬関数の推定法を提案し，EA より優れた方策の推定が可能であることを示している。Syed らの手法は繰り返しマトリクスゲームに対して乗算型重み更新 (multiplicative weights algorithm) を用いて方策の最適化を図っているが，本論文のように sEA 同士がゲーム環境にない問題は対象としていない点で異なっている。

以上の検討を経て，逆強化学習において，最適な行動をとるエキスパートが存在せず理想的な報酬関数を直接獲得できない場合に，不完全な演示しかできない複数の準エキスパートから推定される報酬関数をアンサンブルさせる逆強化学習法を明らにできた。その独自性は，個別の逆強化学習による報酬推定と適応ブースティングを組み合わせた報酬関数の統合により，sE より優れた振る舞いを獲得できることであり，簡便な方法でありながらアンサンブルの効果を逆強化に用いる新たな方法である。

一方，本研究が対象とした Ng らの見習い学習では，個々の sEA から観測される特徴期待値の重み付き平均を用いて求めた報酬関数でも，個々の報酬関数のアンサンブルと同様の性能が得られる可能性がある。計算コストの上では前者が有利となるが，逆強化学習の出力である報酬関数をアンサンブルする

考え方は，特徴期待値を用いない他の逆強化学習やそれらを混合させたアンサンブル学習時にも有効である。今後は，報酬関数のアンサンブルに要する計算コストの評価に加え，本手法が有効に機能する sEA の数と分布および獲得方法の検討，環境の規模・次元に対する計算効率の検討，POMDP 環境における性能評価と性能限界の評価が必要である。

<引用文献>


(1) Andrew Y.Ng and Stuart Russell: “Algorithms for Inverse Reinforcement Learning,” In Proceedings of the 17th International Conference on Machine Learning, pp.663-670 (2000)
(2) 荒井幸代: “逆強化学習によるマルチエージェント系の報酬設定”，計測と制御 Vol.52, No.6, pp.534-539 (2013)
(3) 荒井幸代，堀澤雄介，北里勇樹: “マルチエージェント逆強化学習による報酬設計問題の考察”，人工知能学会全国大会大会論文集 29, pp.1-4 (2015)
(4) Kris Kitani, Brian D. Ziebart, J. Andrew Bagnell, and Martial Hebert: “Activity Forecasting,” Computer Vision – ECCV 2012, Vol.7575 of the series Lecture Notes in Computer Science pp.201-214, Springer (2012)
(5) Y. Freund and R.E.Schapire : “Experiments with a New Boosting Algorithm”, In Proceedings of the 13th International Conference on Machine Learning (1996)
(6) Thomas G. Dietterich: “Ensemble Methods in Machine Learning”, In Proceedings of Multiple Classifier Systems (2000)
(7) 櫻井俊輔，大羽成征，石井信: “学習途上エージェントの挙動に基づく逆強化学習”，信学技報 115(112), pp.95-99 (2015)
(8) Jaedeug. Choi and Kee-Euug Kim: “Inverse Reinforcement Learning in Partially Observable Environments,” The Journal of Machine Learning Research Vol.12, pp.691-730 (2011)
(9) Umar Syed and Robert E. Schapire: “A Game-Theoretic Approach to Apprenticeship Learning,” Advances in Neural Information Processing Systems 20, pp.1449-1456 (2007)


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 5 件）


(1) 冨田真司，濱津文哉，濱上知樹，“準エキスパート集団からのアンサンブル逆強化学習”，電気学会電子情報システム部門論文誌 C，査読有,Vol.137 No.4 pp.667-673,(2017.4) DOI: 10.1541/ieejeiss.137.667 ()
(2) M.Nakata, T.Hamagami, “An Analysis of Rule Deletion Scheme in XCS on Reinforcement Learning Problem”, Journal of Advanced Computational Intelligent Information, 査読有, Vol.21 No.5, pp. 876-884, (2017.5) DOI: 10.20965/jaciii.2017.p0876
(3) Hayato Sasaki, Fumiya Hamatsu, Tomoki Hamagami, “Construction of visual codebook to speed up visual-based simultaneous localization and mapping”, Journal of International Council on Electrical Engineering, 査読有, Vol.7 No.1 (2017.6) 166-173, DOI: 10.1080/22348972.2017.1345371
(4) M.Nakata, T.Hamagami, “Revisit of Rule-Deletion Strategy for XCSAM Classifier System on Classification”, Transaction of Ins. of Sys., Cont. and Eng., 査読有, Vol.30 No.7, pp.273-285, (2017.7) DOI: 10.5687/iscie.30.273
(5) 中尾浩二，外山達斎，林孝則，堀貴雅，濱上知樹，“One-class Support Vector Machine を用いたクラス集合逐次作成法による回転機の異常検知”，電気学会 産業応用部門誌，査読有，Vol.137 No.12 pp.858-865 (2017.12) DOI: 10.1541/ieejias.137.858


〔学会発表〕（計 5 件）


(1) M.Nakata, W.Browne, T. Hamagami, K.Takadama, “Theoretical XCS parameter settings of learning accurate classifiers”, Proc. of GECCO2017, pp. 473-480, (2017.7) DOI: 10.1145/3071178.3071200
(2) X.Li, F.Hamatsu, T.Hamagami, “A Research About Anomaly Detection Method for Multidimensional Time Series Data”, Proc. On The international conference on electrical and engineering, 201702160000009, 1456-1461 (2017.7)
(3) F.Hamatsu, T.Tanaka, T.Hamagami, “Extraction of contours in ultrasound images using a particle filter with limited contour presence range”, The international conference on electrical and engineering, Paper 201703010000001, 2071-2076, (2017.7)
(4) K.Nakao, T.Toyama, T.Hayashi, T.Hamagami, “A rotating machinery fault sign detection method by using multi-level one-class SVMs”, The international conference on electrical and engineering, 201702160000009, 1397-1402, (2017.7)

(5) H.Sasaki, M.Nakata, F.Hamatsu, T.Hamagami, “Effect of Parameter Sharing for Multimodal Deep Autoencoders”, Proc. of IEEE SMC2017, (2017.10) DOI: 10.1109/SMC.2017.8122906

〔図書〕（計 1 件）

(1) 濱上知樹, 機械学習・人工知能　業務活用の手引き～導入の判断・具体的応用とその運用設計事例集～第３章　機械学習とそのアルゴリズム p.47-81 (2017.11)

ISBN- 978-4-86502-142-4

〔産業財産権〕

○出願状況（計　０件）

○取得状況（計　０件）

〔その他〕

http://hamagamilab.ynu.ac.jp

６．研究組織

(1)研究代表者

濱上 知樹(HAMAGAMI, Tomoki)

横浜国立大学・大学院工学研究院・教授

研究者番号：　30334204