人と機械の調和を促進する協調構造学習理論の研究

A study of the collaborative structure learning theory for promoting harmony of man and machine

2157007

研究代表者（助成金受領者）　横浜国立大学大学院　工学府　博士課程後期　恒川裕章
代理　横浜国立大学大学院工学研究院　教授　濱上知樹

［研究の目的］

人と機械が相互に依存し合うことでますます大規模・複雑化するシステムの設計・構築・運用の困難さを克服し，社会の発展サイクルを持続し続けるために，高信頼・安全・堅牢・自律性の源となる新たな協調構造の学習理論を確立する。その核心技術として，人や機械を含む多様な学習主体（エージェント）の集団（マルチエージェント）における「群逆強化学習理論」を研究し，人と機械の調和を促進するための知的相互運用のメカニズムを明らかにする。

［研究の内容，成果］

1）概要

本研究では，目的に掲げた協調知のモデルとして，非エキスパート集団からの見まね学習をアンサンブルによって高度化する「アンサンブル逆強化学習」のアルゴリズムを提案する。Ng らによって提案された逆強化学習の一手法である見習い学習は，エキスパート（EA）と呼ばれるタスク達成可能なエージェントの振る舞いを他の非エキスパートエージェント（sEA）が観測し，未知の報酬空間を推定（逆強化）する。本研究では，この報酬空間の推定を，高次の協調行動や相互干渉を伴う複雑なマルチタスクに応用する。さらに，エキスパートが存在しない場合であっても，複数のエージェント同士の報酬関数をアンサンブルすることで，エキスパートに匹敵する行動が獲得可能な報酬空間を構成できることを示す。この手法を「アンサンブル逆強化学習」と呼ぶ。さらに，不完全知覚の存在する環境においても，アンサンブル逆強化学習は有効であることを明らかにした。

2）逆強化学習とその課題

状態空間 S が k 個の特徴を要素とした特徴ベクトル $\phi: S \rightarrow [0, 1]_k$ を用いて表現可能であるとする。このとき，状態 $s \in S$ に与えられる報酬 $R^*(s)$ は，次式で表される。

$$R^*(s) = w^* \phi(s), w^* \in R^k \cdots\cdots\cdots\cdots\cdots\cdots (1)$$

ここで，w^* は特徴の重みを表す。報酬の最大値を 1 以下に保つために，$\| w^* \|_1 < 1$とする。

方策 ϕ のもとで観測される特徴の期待値を特徴期待値を $\mu(\pi)$ と呼び，次式で定義される。

$$\mu(\pi) = E[\sum_{t=0}^{\infty} \gamma^t \phi(s_t) | \pi] \in R^k \cdots\cdots\cdots\cdots (2)$$

特に，エキスパートの m 個の状態遷移系列 $\{s_0^i, s_1^i \cdots\}_{i=1}^m$ が与えられたとき，エキスパートの特徴期待値 $\mu_E = \mu(\pi_E)$ は次式によって求められる。

$$\mu_E = \lambda_1 \mu(\pi_1) + \lambda_2 \mu(\pi_2) + \cdots + \lambda_n \mu(\pi_n) \cdots\cdots\cdots (3)$$

μ_E を他の特徴期待値の線形和で近似するア

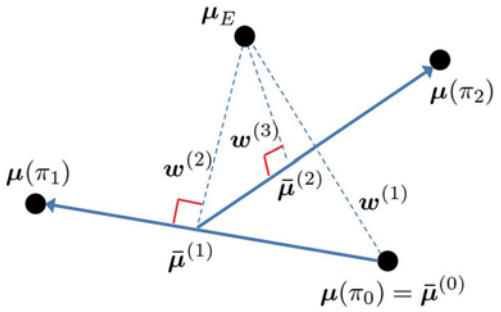

図 1　見まね学習における Projection Method

ルゴリズムの一つが Projection Method（PM）である。PM による $\bar{\mu}^{(i)}$ の導出方法を図 1 に示す。

また PM による方策学習を以下に示す。

1. ランダムに選んだ方策 π_0 のもとで $\mu^0=\mu(\pi_0)$ を計算し，$i=1$ とする。
2. PM によって $\bar{\mu}^{(i)}$，$w^{(i)}$，$t^{(i)}$ をそれぞれ求める。
3. $t^{(i)} \leq \epsilon$ のとき，アルゴリズムを終了する。
4. 報酬関数 $R=w^{(i)}\cdot\phi$ のもとでの最適方策 π_i を強化学習によって求める。
5. 方策 π_i のもとで $\mu^{(i)}=\mu(\pi_i)$ を計算する。
6. $i \leftarrow i+1$ としてステップ 2 に戻る。

環境中に不完全知覚がある場合は，得られた報酬関数をもとにした強化学習は最適な振る舞いとはならず，確率的にタスクが達成できるエージェントとなる。その振る舞いから推定される報酬推定は不完全知覚の影響を受けたものとなる。本研究では，このようなエージェントを，semi-EA（sEA）と呼ぶ。

一方，ノイズの影響を受ける学習の精度を向上させるメタアルゴリズムとして，適応ブースティング（Adaboost）等のアンサンブル学習法が知られている。このメタアルゴリズムを EA が存在しない場合の IRL に応用するために，Adaboost に準じた報酬関数の重み付き統合を行う。すなわち，不完全知覚の影響により最適行動にはなっていない複数の sEA に対して IRL を行い，得られる報酬関数を適切に統合することを試みる。この操作によって，報酬関数に含まれるノイズやランダムな振る舞いを取り除き，不完全知覚の影響を避ける振る舞いの獲得が期待できる。

3）アンサンブル逆強化学習

アンサンブル逆強化学習の概要を図 2 に示す。前述した 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とにする。

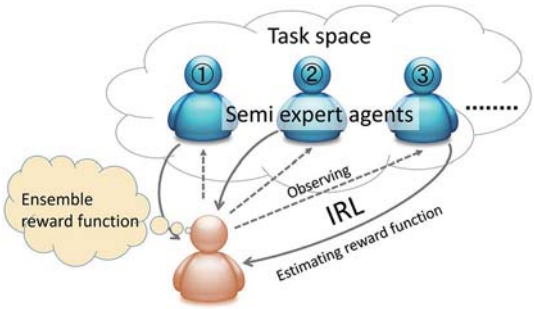

図 2　アンサンブル逆強化学習

信頼度は sEA の振る舞いを表す特徴期待値ベクトル μ 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。

本手法のアルゴリズムの概要を以下に示す。複数の sEA を，$\mathrm{sEA}^{(1)}\cdots\mathrm{sEA}^{(K)}$ と表す。K は sEA の総数である。また，特徴ベクトルの各要素を $l=1, 2\cdots L$ と表す。

1. K 体の sEA から見習い学習を行い，それぞれの特徴期待値と報酬関数を推定する。k=1 とする。
2. $k=1$ から $k=k'$ 番目の sEA の各特徴の分散 σ_l と偏差 $E_1^{(k)}$ を計算する。
3. 各特徴の重要度 $\beta^{(K)}$ と $\mathrm{sEA}^{(K)}$ の信頼度 $\alpha^{(K)}$ を計算する
4. $\beta^{(K)}$ を更新する。

5. $k'=K$ のとき信頼度の正規化後に報酬関数を統合しアルゴリズムを終了する。

図3に，アンサンブル逆強化学習のアルゴリズムを示す。

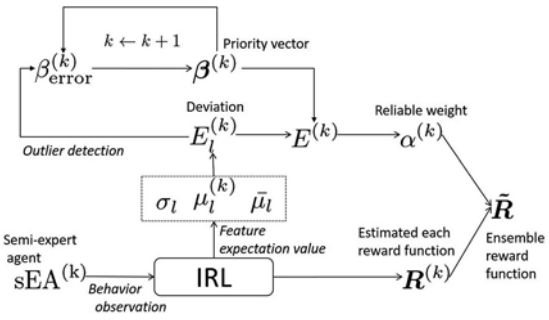

図3　アンサンブル逆強化学習アルゴリズム

4）実験方法

実験に用いる縦18×横18セル（計324セル）のグリッドワールドを図4に示す。エージェントは，左下のスタートから1ステップにつき隣接する周囲の8セルのいずれかに移動しながら右上のゴールまで移動する。

グリッド中には不完全知覚エリア（4種類，計75セル）が存在する。sEA，LAともに，同じ番号がついた不完全知覚セル同士の区別はできない。また，すべてのsEAは不完全知覚セルではランダムな行動をとるものとする。できるだけ少ないステップでゴールに到達するためには，不完全知覚エリアを回避する報酬関数を推測する必要があるが，sEAはいずれも不完全知覚エリアに侵入してしまい，最短での移動が保証されない程度の性能しかない。

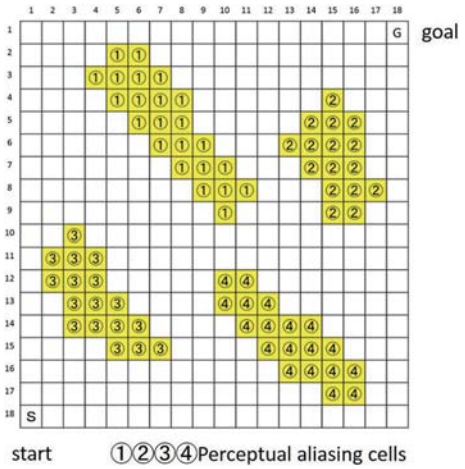

図4　不完全知覚を含むグリッドワールド

本実験では，複数のsEAを，途中で進化を中断させた遺伝的アルゴリズムによって作成する。

各個体の遺伝子は，不完全知覚エリアとゴールを除く248セルにおける行動（8種）と定義し，各ステップにおけるゴールへの接近を評価値とする適応度を用いる。

集団の個体数を100とする。適応度の上位2体をエリートとして保存し，残りの98体をランキング選択と交叉，突然変異を用いて世代更新を行う。

ランキング選択は適応度上位20位のまでの個体を1.8%，21～40位の個体を1.4%，41～60位の個体を1.0%，61～80位の個体を0.6%，81～100位の個体を0.2%の確率で取り出し，これらを2回繰り返すことで交叉に用いる親個体とする。

100体の選択確率の合計は100%である。交叉は遺伝子ごとに20%の確率で交換し，突然変異確率は遺伝子ごとに1％とし，子の2体を次世代の個体とする。すなわち1世代当たりの交叉回数は49回である。進化の過程でエリート個体の適応度が335に達した時点で処理を打ち切り，そのときの遺伝子をsEAの行動ルールとして採用する。このようにして，10体の異なるsEAを作成する。

LAは，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する。LAの強化学習の繰り返し回数を50回，学習率は0.3，割引率は0.99とする。

1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，方策にはϵ-greedy選択を用いる。またランダム行動率ϵは1.0からスタートし，10000ステップごとに0.1減少させ，$\epsilon=0$となってから学習エージェントの特徴期待値を100エピソード分観測する。1エピソードあたりの最大ステップ数は1000である。

5）実験結果

生成した 10 体の sEA を sEA$^{(1)}$～sEA$^{(10)}$とする。それぞれの sEA の観測から，特徴期待値の計測と報酬関数の推定を個別に行う。sEA$^{(1)}$の振る舞いから推定された報酬関数と，sEA の典型的な行動例を図 5 に示す。図より，sEA$^{(1)}$は不完全知覚エリアに侵入をする振る舞いとなることがわかる。ランダム行動により不完全知覚エリアから脱出し，ゴールには到達できるものの，最適な行動にはなっていない。報酬関数をみると，不完全知覚エリアの周辺で価値が上昇していることがわかる。sEA$^{(2)}$以降の振る舞いと報酬関数のも経路中のどこかに周辺より高い価値が推定されており，不完全知覚による冗長な行動が間違った価値関数の推定につながっている。

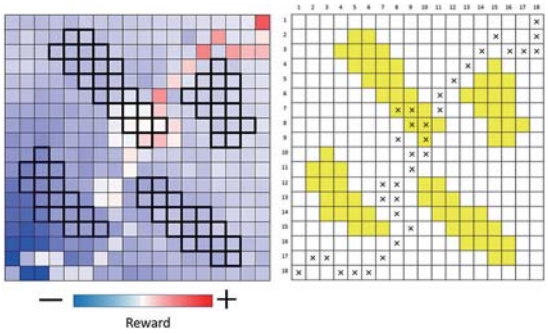

図 5　sEA$^{(1)}$ が獲得した報酬と経路

次に，10 体の sEA から推定した報酬関数の統合結果と，これを用いた LA の学習結果を図 6 に示す。図 6 より，不完全知覚エリア内に与えられる報酬は全て負となり，ゴール付近にのみ正の報酬が現れていることが分かる。以上の結果から，推定された報酬関数を用いて得られる強化学習結果は，ほぼ最適な経路となっていることが明らかとなった。

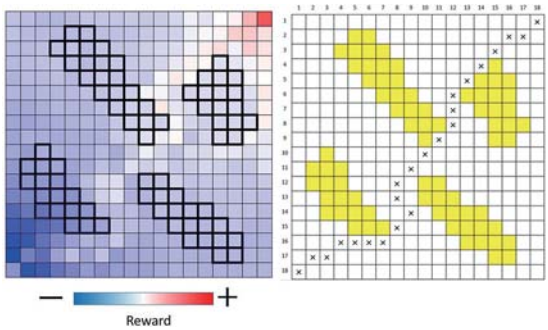

図 6　アンサンブル逆強化による学習結果

6）考察

図 5，6 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ものの，アンサンブル後の報酬関数の方が，ゴールに対して放射状に価値が分布しており，より自然な価値推定になっていることがわかる。

また，アンサンブル前の報酬関数は，経路の途中に局所的な報酬の極大点が存在するものの，アンサンブルされた結果は，スタートからゴールまで単調に価値が増加していることがわかる。さらに，不完全知覚エリアに隣接したセルの価値は経路より低くなっており，不完全知覚エリアを避ける振る舞いが獲得できている。

以上より，アンサンブル後の報酬関数を用いた学習結果は不完全知覚エリアを避けながら確実にゴールに到達する振る舞いが獲得できていることが明らかとなった。

［今後の研究の方向，課題］

アンサンブル逆強化学習は，エージェント間の報酬共有により，協調構造を自律的に学習できる新たな学習理論である。しかし，高次元空間における計算量の理論的解析や，タスクの規模に応じたスケール性については，まだ十分な検討が行われていない。今後はこれらの理論的側面を担保しながら，自動運転タスクへの応用をはかっていく

［成果の発表，論文等］

H. Tsunekawa, T. Suzuki, T. Hamagami, “Examination of skill-based learning by inverse reinforcement learning using evolutionary process”, Intelligent Systems and Control (ISCO), 2015 IEEE 9th International Conference, pp. 1–4 Jan, 2015