1M2-2

報酬関数と状態表現の相互改善による徒弟学習の効率化

Improving Learning Efficiency of Apprenticeship Learning by Mutual Improvement of Reward Function and State Representation

吉永和史 *1　　荒井幸代 *1
Yoshinaga Kazufumi　　Arai Sachiyo

*1千葉大学大学院工学研究科都市環境システムコース
Graduate School of Engineering, Chiba University, Division of Urban Environment Systems

Applying a reinforcement learning framework to the real world problems, shaping of a reward function and construction of a state representation are critical issues to acquire an optimal policy. In the previous studies related to these issues, each issue has been treated independently. In this study, we treat both issues at the same time with enough iterations using an apprenticeship, then realize a mutual improvement. The empirical results show that the mutual improvement makes an appropriate size of state space, and accordingly the learning efficiency becomes increase.

1. はじめに

強化学習を実問題に適用する際,「報酬関数の設計」と「状態空間の設計」がボトルネックである。近年のセンサ性能の向上は，処理すべき状態空間を増大させることにつながり，効率的に学習するには，適切な報酬関数と状態空間の設計が必要となる。

報酬関数の設計については，Ng ら [Ng 00] や Abbeel ら [Abbeel 04] をはじめとした逆強化学習が提案されている。また，状態空間の設計については，基底関数の利用や学習性能に応じた試行錯誤的な手法 [高橋 99] などが提案されている。これらの手法は，いずれも報酬関数と状態空間のうち，一方を所与とし他方を設計する手法である。報酬関数と状態空間は密接に関係しているため，所与とした状態空間または報酬関数に学習性能が依存する。

そこで，本研究では報酬関数または状態空間の一方を所与として他方を設計するのではなく，双方を相互に改善する手法 (以下，相互改善法) を提案する。相互改善法は，エキスパートの行動軌跡を所与とし，この行動軌跡が学習可能な報酬関数と状態空間を設計する。エキスパートの行動軌跡を師，学習エージェントを弟子として，学習エージェントによるエキスパートの行動軌跡の学習を徒弟学習とする。

2. 相互改善法

本研究では，報酬関数の設計と状態空間の設計を相互に繰返し行うアルゴリズムを相互改善法とする。相互改善法の既存アルゴリズムに，Sergey らによる，逆強化学習のための特徴構築 (Feature Construction for Inverse Reinforcement Learning [Sergey 10]; 以下，FIRL) がある。はじめに，Sergey らの相互改善法の有用性を評価するための予備実験 1 を行う。

2.1 FIRL のアルゴリズム

図 1 に FIRL のアルゴリズムを示す。FIRL は，二次計画法を用いて報酬関数を設計する Optimization Step と，回帰木を用いることにより状態空間を設計する Fitting Step から構成される。

FIRL は，最小単位ユニットに細分化された状態空間とそれに対する報酬関数をそれぞれ初期化し，エキスパートの行動軌跡を所与とする。状態空間を，その部分集合である特徴として扱い，アルゴリズムの繰返しごとの分割と報酬を用いた統合によって特徴を更新する。更新された特徴に対して報酬関数を設計する。その後，再び特徴の更新から繰り返すことにより，適切な状態空間とそれに合う報酬関数を設計する。

連絡先: 吉永和史，千葉大学大学院工学研究科都市環境システムコース，千葉市稲毛区弥生町 1-33, k.yoshina0318@gmail.com

エキスパートの行動軌跡 D を所与
0. 初期化
　繰り返し回数：$i \leftarrow 1$
　状態空間：$S \leftarrow$ 最小単位ユニット
　報酬関数：$R \leftarrow D$ を学習可能な報酬関数
　特徴集合：$\Phi \leftarrow$ 全単位ユニットを同一の特徴
　行列 $T_{R\to\Phi}$，$T_{\Phi\to R}$，状態価値 V を任意に初期化
以下を十分な回数繰り返し：
1. Optimization Step
　行列 $T_{R\to\Phi}$，$T_{\Phi\to R}$，状態価値 V を更新
　二次計画法により R を更新
2. Fitting Step
　Ⅰ) 特徴 ϕ_α，ϕ_β が次の条件を満たすとき統合
　　条件 A：$\theta_{s_{\alpha_1} a s_{\beta_1}} \neq 0$　　$s_{\alpha_1} \in \phi_\alpha, s_{\beta_1} \in \phi_\beta$
　　条件 B：$r_\alpha = r_\beta$
　Ⅱ) 回帰木を操作し，Φ を更新
　Ⅲ) 各 ϕ ($\phi \in \Phi$) について：
　　特徴内の報酬を平均化した報酬関数 $\hat{R}$ を作成
　　$\hat{R}$ を用いて価値反復を行い方策を獲得
　　方策が D に一致する場合，$R \leftarrow \hat{R}$ とし，以後その特徴を更新しない
3. $i \leftarrow i+1$ とし，1. へ

図 1: FIRL のアルゴリズム

2.2 予備実験 1 (FIRL の評価)

2.2.1 実験設定

■実験環境

実験には図 2 に示す 8×8 の grid world を用いる。FIRL における初期の状態数は 64，初期報酬は Abbeel の逆強化学習で求め，行動は { 上，右，下，左 } の 4 つとする。また，右下の状態 $(7,7)$ を初期状態，左上の状態 $(0,0)$ を終端状態とし，エキスパートの行動軌跡は,「壁沿い」と「対角線」の 2 通りとする。図 2 より，壁沿いは決定的方策をもつ状態を経由し，対角線は確率的方策をもつ状態を経由する。

■評価方法

性能評価の指標として学習効率と状態数を用いる。学習効率は「最短経路獲得に要するエピソード数」とする。エピソードとは初期状態から行動選択を繰り返し，学習をしながら終端状態へ到達するまでの「状態-行動系列」をさす。本研究では，最短経路獲得までに要するエピソード数を学習効率の指標とする。状態数はアルゴリズムの終了時点で設計された状態の数である。

これら 2 つの指標について，ともにエキスパートの行動軌跡を所与とする，Abbeel の逆強化学習 [Abbeel 04] と FIRL

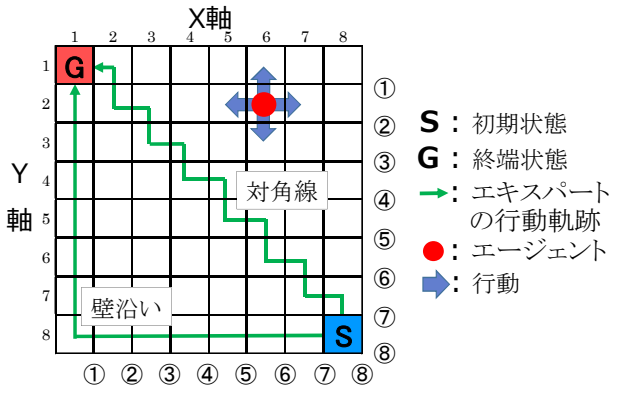
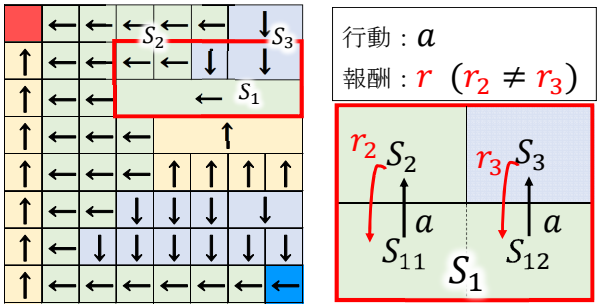


図 2: 実験環境 (grid world)



(a) 設計された状態空間　(b) 不完全知覚 (図 3(a) の赤枠内)

図 3: FIRL によって設計された状態空間と不完全知覚

表 1: 予備実験 1 ：最短経路獲得に要する 10 試行平均のエピソード数と状態数

	手法	エピソード数 (標準偏差)	状態数 (標準偏差)
壁沿い	逆強化学習	2076.9 (65.6)	64.0 (0.0)
	FIRL	2198.9 (142.5)	55.3 (4.3)
対角線	逆強化学習	2425.1 (109.1)	64.0 (0.0)
	FIRL	2254.8 (167.0)	53.0 (0.0)

を比較する。

■分割条件

FIRL を用いるにあたり，特徴の分割条件はアルゴリズムの繰り返し回数 i を基準にする。ある特徴に含まれる状態 $s_{x,\,y}$ について，$x \geq i$ と $x < i$，また $y \geq i$ と $y < i$ の組み合わせによって，1 つの特徴が最大 4 つの特徴に分割される。

図 1 のアルゴリズムより，全状態を一つの特徴として初期化し，特徴分割の順番は図 2 における ① → ② → ③ → ··· → ⑦ の順番で分割する。

2.2.2 実験結果

2 種類のエキスパートの行動軌跡について，最短経路を獲得するまでに要したエピソード数の 10 試行平均と標準偏差，設計された状態空間の状態数と標準偏差を表 1 に示す。最短経路獲得までのエピソード数の差の検証を，有意水準 1% の t 検定で行った。その結果，壁沿いについては有意差がみられ Abbeel の逆強化学習の学習効率がよく，対角線については有意差がみられなかった。状態数については壁沿いが約 13.6%，対角線が約 17.2% 減少した。また，これらの実験設定の下で獲得した方策は，すべての実験においてエキスパートの行動軌跡が再現された。

2.2.3 考察

壁沿いの行動軌跡を用いた際に，FIRL により設計された状態空間とそこで学習された行動を図 3(a) に示す。壁沿いの行動軌跡において FIRL の学習効率が落ちた要因として，図 3(a) の状態 s_1 において，図 3(b) のような不完全知覚が発生したことが挙げられる。s_1 において，行動をランダム選択で学習している最中に「上」の行動を選択した場合，4 つの状態のいずれかに遷移する。遷移先の状態によって異なる報酬が得られ，学習の収束に時間を有したと考えられる。また，FIRL によって設計された状態空間ではさらに状態を統合できる箇所が観察された。

予備実験 1 により，相互改善法の FIRL は，学習効率において Abbeel の逆強化学習に同等か劣る結果となり，状態数は行動軌跡によらず減らすことができた。よって，状態数の削減という点では FIRL は有用といえる。次章で，予備実験 1 において課題と考えられる学習効率を上げ，状態の統合が促進する手法を提案する。

3. 提案手法

FIRL のアルゴリズムを二点変更した手法を提案手法とする。

■一点目：行動の考慮

予備実験 1 で，学習効率の低下は状態を統合することによる不完全知覚の発生が原因であった。不完全知覚の原因として，FIRL において特徴を統合する際の条件が挙げられる。これは，図 1 の 2.Ⅰ) に示した部分で，統合の条件は「2 つの特徴が隣接している，かつ二つの特徴の報酬が等しいとき」である。FIRL では特徴の統合に，報酬のみを考慮しているため不完全知覚が発生したと考えられる。そこで，提案手法では特徴の統合を行う際に「2 つの特徴で学習される行動が等しいとき」という，行動を考慮した条件を追加する。

■二点目：分割停止条件の変更

予備実験 1 では，設計された状態空間にさらに状態を統合できる箇所が観察された。そこで，FIRL において一つの特徴がより多くの状態を含めるよう，条件を緩和する。

図 1 の 2.Ⅲ) に示した部分に注目する。特徴 ϕ の分割を停止するための条件は,「ϕ に含まれる状態の報酬を平均化し，平均化した状態空間における獲得方策が，所与の行動軌跡に一致するとき」である。これは，獲得方策が初期状態から終端状態の全状態行動対で，所与の行動軌跡に一致している必要がある。

ここで，最終的な状態数を削減するための分割停止条件の緩和として,「獲得方策と所与の行動軌跡の部分一致」を提案する。図 1 の 2.Ⅲ) では，ある特徴 ϕ を指定して操作を行うため,「獲得方策が ϕ において所与の行動軌跡に一致」とできる。これにより，行動を維持しつつ状態数を減らすことが可能と考えられる。この条件緩和において，特徴が行動軌跡に含まれない場合を考慮する必要がある。この際の行動軌跡をもたない特徴に対する操作として考えられるのが,「分割の続行」と「分割の停止」である。初期の条件も含め図 4 の 3 つの条件を予備実験 2 として実験し，最も性能のよい条件を選択する。

● 予備実験 2　結果と考察

2.2 節の実験環境に従い図 4 の条件で実験した結果を表 2 に示す。条件 ① は，分割停止条件の緩和前のため，他の 2 条件と比べ状態数が大きくなっている。条件 ② は，行動軌跡によらず状態数が最も小さくなったが，最短経路学習までのエピソード数は壁沿いにおいて 2 番目，対角線については最も大きくなった。原因として，条件 ② は行動軌跡がない状態について，その時点で分割をやめるため不完全知覚が発生していると考えられる。よって，提案手法の分割停止条件は，図 4 の「条件 ③　獲得方策が状態内で一致，行動軌跡がない状態は分

① 獲得方策が状態空間全体で一致

② 獲得方策が状態内で一致，行動軌跡がない状態は分割停止

③ 獲得方策が状態内で一致，行動軌跡がない状態は分割続行

図 4: 予備実験 2 ：分割停止条件の種類

表 2: 予備実験 2 ：最短経路獲得に要する 10 試行平均のエピソード数と状態数　条件は図 4 に対応

	条件	エピソード数 (標準偏差)	状態数 (標準偏差)
壁沿い	①	1584.0 (1146.8)	24.8 (16.5)
	②	68.5 (61.4)	12.3 (0.5)
	③	45.6 (23.4)	14.4 (1.3)
対角線	①	2440.0 (100.8)	58.0 (0.0)
	②	3269.1 (573.7)	27.0 (1.1)
	③	2693.2 (109.6)	58.7(2.9)

割続行」を用いる。

4. 計算機実験

提案手法の評価実験を，2.2 節の実験環境に従い行う。

4.1 実験 1：提案手法の性能評価

所与にエキスパートの行動軌跡を用いる，Abbeel の逆強化学習，FIRL，提案手法を比較する。FIRL と提案手法の初期報酬には Abbeel の逆強化学習によって求めた報酬を用いる。

■実験 1：結果

表 3 に実験結果を示す。壁沿いの行動軌跡において，提案手法は最短経路獲得までのエピソード数，状態数ともに他の手法に対して優れた結果が得られた。しかし，対角線の行動軌跡において，提案手法の最短経路獲得までのエピソード数は，他の手法に対して有意水準 1%の t 検定において有意差があると判断でき，劣っていた。状態数は減少しているが FIRL の減少数の方が大きい。設計された報酬関数と状態空間において学習を行った結果，全ての試行において行動軌跡が学習された。

■実験 1：考察

提案手法において壁沿いの行動軌跡を所与とした場合に，状態数が減少した原因として，行動軌跡上の連続する状態において同じ行動を示すことが挙げられる。このため，分割が進む前に緩和した分割停止条件に当てはまり，最終的な状態数が減少したと考えられる。また，対角線の行動軌跡を所与とした場合，行動軌跡が連続する状態において異なるため FIRL と同様に分割され，かつ，統合条件に「行動の一致」が追加されたことにより，FIRL に比べ状態が統合されなかった。ここでの，提案手法のエピソード数の増加は，状態数の増加が原因と考えられる。

4.2 実験 2：初期報酬による性能評価

FIRL と提案手法について，初期報酬による影響を調べるため，Abbeel の逆強化学習で得られた報酬関数を用いた場合と，終端状態にのみ報酬を置いた場合の比較を行う。評価には，エピソード数，状態数に加え，軌跡学習数を用いる。ここでの軌跡学習数は，10 試行中に学習された方策が所与としたエキスパートの行動軌跡に一致した回数とする。

表 3: 実験 1：最短経路獲得に要する 10 試行平均のエピソード数と状態数

	手法	エピソード数 (標準偏差)	状態数 (標準偏差)
壁沿い	逆強化学習	2076.9 (65.6)	64.0 (0.0)
	FIRL	2198.9 (142.5)	55.3 (4.3)
	提案手法	45.6 (23.4)	14.4 (1.3)
対角線	逆強化学習	2425.1 (109.1)	64.0 (0.0)
	FIRL	2254.8 (167.0)	53.0 (0.0)
	提案手法	2693.2 (109.6)	58.7 (2.9)

表 4: 実験 2：最短経路獲得に要する 10 試行平均のエピソード数と状態数の初期報酬による差

	手法	エピソード数 (標準偏差)	状態数 (標準偏差)	軌跡 学習数
壁沿い 逆強化	FIRL	2198.9 (142.5)	55.3 (4.3)	10
	提案手法	45.6 (23.4)	14.4 (1.3)	10
壁沿い 終端状態	FIRL	468.4 (49.6)	58.8 (1.8)	2
	提案手法	244.2 (86.2)	39.1 (1.4)	10
対角線 逆強化	FIRL	2254.8 (167.0)	53.0 (0.0)	10
	提案手法	2693.2 (109.6)	58.7 (2.9)	10
対角線 終端状態	FIRL	462.5 (15.1)	64.0 (0.0)	0
	提案手法	662.2 (103.1)	54.7 (0.7)	0

■実験 2：結果

表 4 に実験結果を示す。最短経路獲得に要するエピソード数は，エキスパートの行動軌跡が学習できたかによる違いはみられず，提案手法の壁沿いの行動軌跡を除き，最短経路獲得に要するエピソード数は終端状態にのみに報酬を置いた場合に減少した。状態数は，提案手法の対角線の行動軌跡の場合を除き，終端状態のみに報酬を置いた方が増加した。終端状態にのみ報酬を置いた際の行動軌跡学習数は，壁沿いの行動軌跡の場合に提案手法は 10 試行全て成功しており，FIRL は 2 回の成功であった。また，対角線の行動軌跡の場合は両手法とも行動軌跡を学習できなかった。

■実験 2：考察

初期報酬を終端状態のみにした場合に，状態数が増加した原因として，アルゴリズム中の特徴の統合条件が挙げられる。統合条件には「報酬が等しい」という条件が含まれており，逆強化学習による報酬を初期報酬とした場合と比較して，隣接する特徴同士の報酬の差が大きくなったため，この条件に適合しなくなった。初期報酬を終端状態のみに設定した場合，両手法において対角線の行動軌跡が学習されない理由として，grid world の性質が挙げられる。grid world では，行動選択において「上」または「左」を選択すると等しく終端状態に近づく。初期報酬が終端状態のみの場合，各状態において「上」または「左」のどちらが最適か示されず，対角線の行動軌跡のように確率的方策をもつ状態を経由する場合，エキスパートの行動軌跡を学習することが困難と考えられる。

4.3 実験 3：分割の順番に性能評価

提案手法において分割の順番の影響を調べるため，分割の順番を変え実験する。これまでの，予備実験，実験 1，実験 2 では図 5 Patt.1 の順番で分割した。しかし，分割停止の条件から，分割の順番は設計される状態空間に影響を与えると考えられる。そこで，実験 3 では図 5 示す順に分割の順番を変更し，設計された状態空間での状態数と学習効率，軌跡学習数を比較する。初期報酬は，Abbeel の逆強化学習によって求める。

■実験 3：結果

表 5 より，両行動軌跡ともに Patt.1 から分割順を変えると，最短経路獲得に要するエピソード数が増加した。状態数につい

番号は図 2 に対応

Patt. 1 ① → ② → ③ → ··· → ⑦

Patt. 2 ⑦ → ⑥ → ⑤ → ··· → ①

Patt. 3 ④ → ⑥ → ② → ⑤ → ③ → ⑦ → ①

図 5: 実験 3：分割順

表 5: 実験 3：最短経路獲得に要する 10 試行平均のエピソード数と状態数の分割順による差

	分割順	平均エピソード数 (標準偏差)	状態数 (標準偏差)	軌跡 学習数
壁沿い	Patt.1	45.6 (23.4)	14.4 (1.3)	10
	Patt.2	8117.8 (1326.6)	52.5 (1.4)	7
	Patt.3	4941.0 (847.9)	37.0 (5.7)	0
対角線	Patt.1	2693.2 (109.6)	58.7 (2.9)	10
	Patt.2	3134.9 (1090.2)	59.1 (5.2)	4
	Patt.3	3857.1 (1079.7)	57.9 (6.0)	3

ても，分割順を変えることにより壁沿いの行動軌跡で差が生じた。軌跡学習数は，壁沿いの Patt.2 のみ行動軌跡を学習可能な状態が設計されず，所与の行動軌跡によらず分割順を Patt.1 以外にした場合に軌跡学習数が下がった。

■実験 3：考察

実験 3 の結果より，提案手法においては分割の順番により学習効率に大きな差が生じる。このことから，設計者に知識がある環境の場合，知識を利用することで，より学習に適した状態空間を設計することができる。しかし，状態空間の設計は未知の状態空間に対して行われるため，今後の課題として未知の状態空間を設計する場合の分割の順番を考慮する必要がある。

4.4 まとめ

実験 1 より，提案手法は Abbeel の逆強化学習や FIRL と比較し，エキスパートの行動軌跡が決定的方策をもつ状態を経由する壁沿いの場合，エキスパートの行動軌跡を効率よく学習可能な状態空間が設計された。一方，エキスパートの行動軌跡が確率的方策をもつ状態を経由する対角線の場合，FIRL と比較し状態数の多い状態空間が設計され，学習効率が下がった。よって，確率的方策をもつ状態を含む場合の状態空間の設計条件を，検討する必要がある。

実験 2 において，初期報酬で報酬を終端状態にのみ設定した場合，提案手法は壁沿いの行動軌跡を所与とした全試行において行動軌跡を学習できた。一方，対角線の行動軌跡を所与とした場合，行動軌跡の学習はできなかった。これは，毎ステップで行動が変わる行動軌跡を，終端状態のみの初期報酬で正確に学習することが困難であったためと考えられる。

実験 3 より，提案手法における状態設計は分割の順番に依存することが分かる。これにより，設計者の経験を活かした状態空間の設計が可能であるが，未知の環境については，分割の順番を考慮する必要がある。

5. 結論および今後の課題

本研究では，強化学習を実問題に適用する際のボトルネックである「報酬関数の設計」と「状態空間の設計」の問題に着目した。それぞれの問題に対して，報酬関数または状態空間の一方を所与としたアルゴリズムの提案はあるが，両者を相互に改善することが可能な相互改善法に注目した。

相互改善法の一つである FIRL を用いた場合，不完全知覚が発生し，また，設計された状態空間の中にはさらに統合可能な状態が観察された。そこで，FIRL の状態設計条件を変更した手法を提案した。計算機実験の結果，提案手法はエキスパートの行動軌跡が決定的方策をもつ状態を経由する場合に状態数を減少させ，高い学習効率を示す状態空間の設計を可能にした。一方，確率的方策をもつ状態を経由する場合に，FIRL よりも学習性能が下がった。このことから，最適な行動が複数存在する状態に対して，エキスパートの行動軌跡が唯一与えられるときの状態空間の設計法が必要である。

よって，複数の決定的方策を有する場合の，学習に適した状態空間の設計が必要である。さらに，FIRL，提案手法ともに初期報酬の影響を受けるため初期報酬の設計方法，また，アルゴリズム中での特徴の分割順の考慮を今後の課題とする。

参考文献

[Ng 00] A. Ng, and S. Russell: Algorithms for inverse reinforcement learning, Proceedings of the 17th International Conference on Machine Learning, pp.663–670, 2000.

[Abbeel 04] P. Abbeel, and A. Ng: Apprenticeship learning via inverse reinforcement learning, Proceedings of the 21th International Conference on Machine Learning, ICML ’04, 2004.

[高橋 99] 高橋 泰岳, 浅田 稔: 実ロボットによる行動学習のための状態空間の漸次的構成, 日本ロボット学会誌, Vol. 17, No. 1, pp. 118-124, 1999.

[Sergey 10] S. Levine, Z. Popović and V. Koltun: Feature construction for inverse reinforcement learning, Proceedings of the 24th Neural Information Processing Systems, 2010.