3P-02

強化学習における報酬関数と状態空間表現の相互改善法の評価

吉永　和史†　　荒井　幸代‡

千葉大学工学部都市環境システム学科†　　千葉大学大学院工学研究科‡

1. はじめに

強化学習は，人工知能研究における計画型問題へのアプローチとして期待されている。しかし，所与とする「報酬関数の設計」と「状態空間の構築」が実問題の適用におけるボトルネックである。これらの問題に対するアプローチとして，報酬関数の設計については，Ng ら[1]や Abbeel ら[2]をはじめとした逆強化学習が提案されている。また，状態空間の構築については，基底関数の利用[3]や学習性能に応じた試行錯誤的な手法[4]などが提案されている。これらのアプローチは，報酬関数の設計には状態空間を，状態空間の構築には報酬関数を，それぞれ一方を所与とした方法である。報酬関数と状態空間は密接に関係しているため，所与とされる状態空間または報酬関数に学習性能は依存する。

そこで，本研究では一方を所与とし他方を設計するのではなく，双方(報酬関数と状態空間)を相互に改善する手法(以後，相互改善手法)を紹介し，計算機実験の結果から相互改善法を学習効率と構築された状態空間の状態数について評価を示す。

2. 相互改善法

本研究では，Sergey ら[5]による，Feature Construction for Inverse Reinforcement Learning (以後，FIRL)を報酬関数と状態空間表現の相互改善法の一つとして位置づける。

図 1 に FIRL のアルゴリズムを示す。FIRL のアルゴリズムは大きく 2 つに分かれており，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構築する Fitting Step に分けられる。

FIRL は，ある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初期状態として与え，エキスパートの行動を所与とする。状態空間に対して報酬の大きさによって状態の統合，分割を行い，適切な状態空間の構築とそれに合った報酬関数を設計する。

1. 繰り返し回数：$t = 0$ とし，報酬関数と状態空間を最小ユニット単位に初期化。エキスパートの行動を所与とする。

以下を十分に繰り返す：

Optimization Step

2. 所与または前回の繰り返しで得られた報酬関数，状態空間をもとに二次計画法を解き新しい報酬関数を得る。

Fitting Step

3. 各状態において，次の状態統合条件をともに満たす状態同士を統合する。
   ・隣接している　　・同じ報酬である
4. 回帰木を分割することにより，状態空間を再構築する。
5. 以下各状態について繰り返す：
   I. 状態内の報酬を平均化する。
   II. 価値反復により方策を得る。
   III. 得られた方策がエキスパートの方策に従う場合，今後分割しない。
6. $t \leftarrow t + 1$ とし，2. へ戻る。

図 1　FIRL のアルゴリズム

3. 計算機実験

■実験環境

報酬関数と状態空間表現の相互改善法の評価を行う。実験 1 で，エキスパートの行動軌跡から報酬関数を求める方法である Abbeel の逆強化学習と FIRL の比較を行う。

実験は，図 2 に示す 8×8 の Grid World 環境において行う。エージェントの行動は〈上，右，下，左〉の 4 つ，エキスパートの行動軌跡は，時刻ごとの行動の変更回数が少ない「右回り」と，多い「対角線」の 2 通りとする。FIRL にお

Evaluating Mutual Improvement of Reward Function and State Representation on Reinforcement Learning Approach
†Kazufumi Yoshinaga ‡Sachiyo Arai
†Faculty of Engineering, Chiba University
‡Graduate School of Engineering, Chiba University

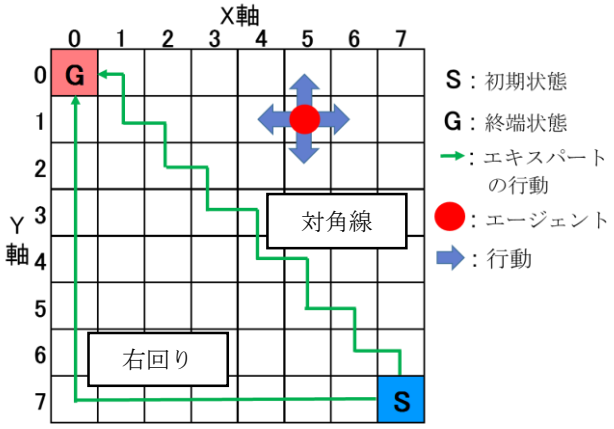

図 2　実験環境(Grid World)

ける初期の状態分割数を 64 とし，各状態に対する報酬は Abbeel の逆強化学習により得られた報酬を用いる。

■実験結果

各手法の学習効率を比較する。学習効率は，最短経路を獲得するまでにかかるエピソード数を指標とする。表 1 に各手法における，2 種類のエキスパートの行動軌跡について，最短経路獲得に要したエピソード数の 10 試行の平均と標準偏差，状態数をまとめる。

最短経路獲得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，右回りについては有意差がみられ IRL 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。右回りにおいて FIRL の学習効率が落ちた要因として，行動軌跡上でない状態で状態統合が進んだことが挙げられる。これにより，終端状態までに通過する状態数は，行動軌跡外を通った方が少なくなり，行動軌跡の学習に時間を有したためと考えられる。標準偏差については，右回り，対角線ともに FIRL が大きくなっている。これは，状態の統合により状態遷移が複雑化し，探索の順番により学習に差が生じるためと考えられる。状態数は平均で，右回りが約 13.6%，対角線が約 17.2%減少した。また，獲得した方策は各手法においてエキスパートの行動軌跡が学習された。

4. おわりに

本研究では，報酬関数と状態空間表現の相互改善法として，FIRL を紹介し，計算機実験による評価を行った。

計算機実験より，Abbeel の逆強化学習の報酬を初期値に用いた場合，学習効率の面では行動の変更回数が少ない右回りの学習において IRL

表 1 最短経路獲得に要するエピソード数と状態数

	手法	平均	標準偏差	平均状態数
右回り	IRL	2076.9	65.6	64.0
	FIRL	2198.8	142.5	55.3
対角線	IRL	2425.1	109.1	64.0
	FIRL	2254.8	167.0	53.0

に劣る結果となった。一方，行動の変更回数が多い対角線の行動軌跡に関しては IRL と同等な結果が得られた。また，エキスパートの行動軌跡によらず状態数は削減されており，この点において FIRL は有用といえる。しかし，構築された状態空間には，さらに統合が可能と判断できる状態がみられた。

したがって，今後の課題として，さらに状態数の少ない状態空間とそれに合った報酬関数の設計が挙げられる。その方法として状態統合条件の変更が考えられる。FIRL において状態の統合が行われるのは，統合する状態同士が「隣接している」かつ「同じ報酬である」ときである。「同じ報酬」から「近い報酬」とする条件の緩和も考えられるが，この状態統合条件は報酬の大きさのみを考慮している。そのため，状態統合において，本来異なる行動を示す 2 つの状態が統合され，不完全知覚が発生すると考えられる。これを解決するために，状態統合条件においてエージェントの行動を考慮する必要がある。

参考文献

[1] Andrew Y. Ng, and Stuart J. Russell. “Algorithms for Inverse Reinforcement Learning” *In Proceedings of the 17th International Conference on Machine Learning*, 2000, pp.663-670.

[2] Pieter Abbeel, and Andrew Y. Ng. “Apprenticeship learning via inverse reinforcement learning” *Proceedings of the 21th International Conference on Machine Learning, ACM*, 2004.

[3] Richard S. Sutton and Andrew G. Barto. “Reinforcement learning: An introduction” A Bradford Book, The MIT Press, 1998.

[4] 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, 1999, pp.118-124.

[5] Sergey Levine, Zoran Popović and Vladlen Koltun. “Feature construction for inverse reinforcement learning” *Proceedings of the 24th Neural Information Processing Systems*, 2010.