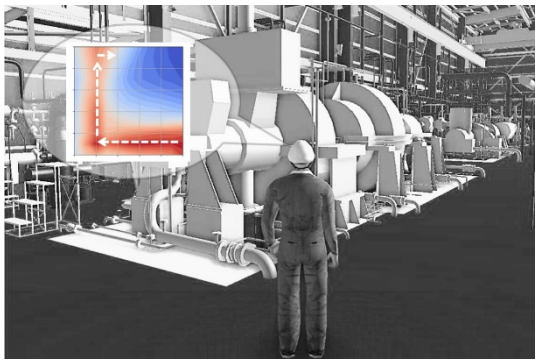
熟練者の操作習得とノウハウ可視化に寄与する逆強化学習技術

Inverse Reinforcement Learning Technology Contributing to Imitation and Know-How Visualization of Plant Expert Operations



中川 陽介*1 Yosuke Nakagawa
小野 仁意*2 Hitoi Ono
筈井 祐介*1 Yusuke Hazui
荒井 幸代*3 Sachiyo Arai

プラントの非定常手動操作は，運転状態に応じた適切な判断が求められ，運転員の知見や技量に大きく依存している。深層強化学習は，状態を認識し，その状態に対する評価指標に基づき一連の操作を学習でき，これまで暗黙知とされてきた操作指針の明文化に繋がる手法として期待されている。しかし，学習に必要な評価指標の設計が難しいことや，学習結果の説明性等の課題があり，安全性・信頼性が求められるプラント操作への適用が進んでいなかった。

本報では，千葉大学との共同研究にて開発した逆強化学習技術を用いて熟練者の操作を習得するとともに，学習結果の可視化を通じてそのノウハウを推量する技術について紹介する。

1. はじめに

プラントにおける起動・停止などの非定常手動操作は，プラント状態に応じて操作のタイミングやその量を変える必要があり，運転員の知見や技量に大きく依存している。従来は，運転員によるプラントへの影響を低減するため，熟練者の操作手順をプラント状態に応じて if-then のルール形式で表し，自動化や経験の浅い運転を補佐する運転システムを構築することが一般的であった。しかし，熟練者の操作には暗黙知が含まれることが多く，これをルール形式で表現することは容易ではない。

近年，計算機スピードの向上に伴い状況認識を担う深層学習の発展により，状況に応じた操作（行動）の習得を担う強化学習と組合せた深層強化学習が，一連の操作，例えば起動開始から完了までの評価指標（報酬）を最大化することで，適切な行動を習得できる方法として期待されている。AlphaZero[(1)]に代表される対戦ゲームの分野においては，深層強化学習により人を上回るパフォーマンスが達成可能であることが示され，プラント操作の分野においても適用[(2)]が期待されている。

一方で，熟練者の操作を習得するための報酬設計が容易ではないこと，得られた学習結果に説明可能性が求められること，またより広範な運転条件にも適用可能なロバスト性のある操作指針の習得等に課題がある。

本報では，これらの課題に対し，教師とする熟練者の操作データから報酬を推定するとともに，熟練者の操作を習得できる逆強化学習技術を開発し，蒸気による配管の暖気操作を例にシミュレーションにより検証した。その結果，熟練者の操作手順を習得でき，また，得られた報酬の可視化により，熟練者のノウハウを推量した。

*1 ICT ソリューション本部 CIS 部
*2 ICT ソリューション本部 CIS 部 主席技師 工博
*3 千葉大学 教授 工博

以降，2章では，逆強化学習の概要を示し，3章では蒸気配管の暖気操作での学習結果や可視化結果について述べ，4章でまとめを述べる

2. 逆強化学習技術

2.1 強化学習と逆強化学習

強化学習は機械学習手法の1つで，試行錯誤を通じて環境から得られる報酬（評価指標）を最大化するような操作手順を学習するもので，個々の操作に対する報酬の最大化ではなく，一連の操作における累積報酬を最大化する。強化学習をプラント運転操作に適用するに当たり，解決したい課題における報酬設計が容易ではないこと，プラント設計者や運用者が想定し得ない新規な操作手順は計算に現れない新たなリスクが潜んでいる可能性があること，得られた学習結果や操作手順についての説明性を求められること等の課題がある。

これらの課題に対し，熟練者の操作データを用いることで，操作手順の習得とともに，報酬の推定も可能な技術として，逆強化学習がある。熟練者の操作を単に真似るだけであれば，古くから模倣学習などがある。これらに対し逆強化学習は，推定した報酬を学習時とは異なる運転条件に転移し，これを用いて新たに強化学習を行うことにより，報酬設計をすることなく熟練者の操作データがない運転条件においても学習できる。その結果，少ないデータや試行回数で広範な運転条件にも適用可能な操作手順を学習できる可能性がある。

2.2 開発した逆強化学習技術

本報では，従来の逆強化学習に比べて，より複雑な行動を習得できるとされ，深層学習の技術である敵対的生成ネットワークを取り入れた逆強化学習の１手法である AIRL（Adversarial Inverse Reinforcement Learning）[(3)]を用いた。AIRL の構成を図1に示す。

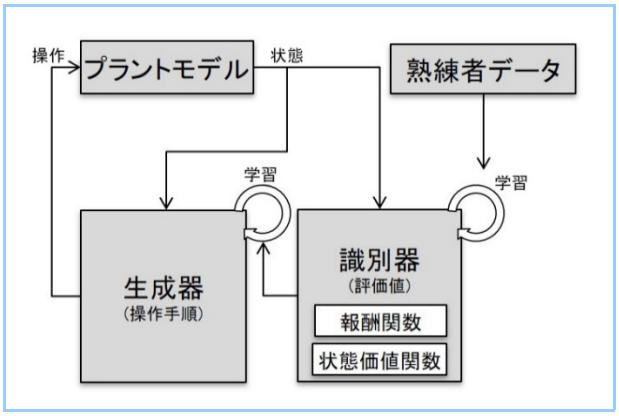


図1 AIRL の構成図
敵対的逆強化学習である AIRL の学習器の構成を示す

敵対的逆強化学習は，2つのニューラルネットワーク（生成器と識別器）で構成される。生成器では，強化学習により操作手順を学習し，その手順をプラントに入力し操作データ（状態変化履歴）を生成する。識別器では，学習により生成された操作データ（贋）か，熟練者の操作データ（真）かを識別する。生成器は識別器を騙そうと，より熟練者に近い操作手順を学習するのに対し，識別器は生成器に騙されないように真贋をより見分けられるように学習する。生成器が熟練者の操作を学習することができれば，識別器は生成器による操作と熟練者の操作を分類できなくなり，結果として生成器が熟練者の行動を模倣する学習が出来たことになる。

AIRL の識別器では，運転操作の評価指標を報酬関数と状態価値関数の2つに分けて学習する。報酬関数は各状態における操作良しあしを相対的に表し，状態価値関数は初期状態からの累積報酬を表し，最終目標を示す。これらの関数を可視化することで，熟練者のノウハウを推量でき，学習結果の説明に繋げることができる。

一方で，逆強化学習で得られる操作手順は，教師データとして与えた熟練者操作近傍の限ら

れた範囲しか習得できない。より広範囲な運転条件に適用するためには，新たな条件下での再学習が必要となる。本報では，AIRL の機能拡張を図り，複数の運転条件を交互に(同時に)学習する同時学習方法を開発した。構成図を図2に示す。

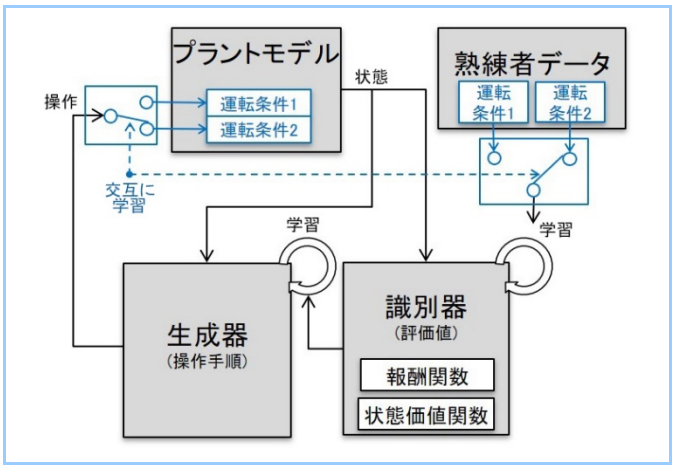


図2 開発した逆強化学習器の構成
複数運転条件を交互に学習するよう従来の AIRL の機能拡張した構成を示す

3. プラントシミュレータでの開発技術の検証

3.1 対象操作の概要

蒸気配管の暖気操作の概略を図3に示す。この操作は，ベント弁，入口弁を用い，上流プラントから供給される高温高圧の蒸気により，配管内を大気状態から供給元の圧力，温度に近い状態まで少ない蒸気消費量で昇温，昇圧するもので，完了後の蒸気は，タービンなどに供給される。操作に際し，タービン翼損傷や配管内でのハンマリングなどのトラブルを防止するため，配管内の蒸気がミストなどの液滴を含まないよう過熱状態を保つ必要がある。そのため，熟練者は①ベント弁全開の状態のまま入口弁を微開(昇温)，②ベント弁を微小開度まで閉とした後に入口弁を全開(昇圧)，③入口弁全開の状態のままベント弁を中間開度まで開(再昇温)の段階的操作を手動で行っている。

さらに熟練者は，大気温度による放熱量の差異を考慮し，弁操作量や操作タイミングを変えてより少ない蒸気消費量での昇温，昇圧操作を行っている。

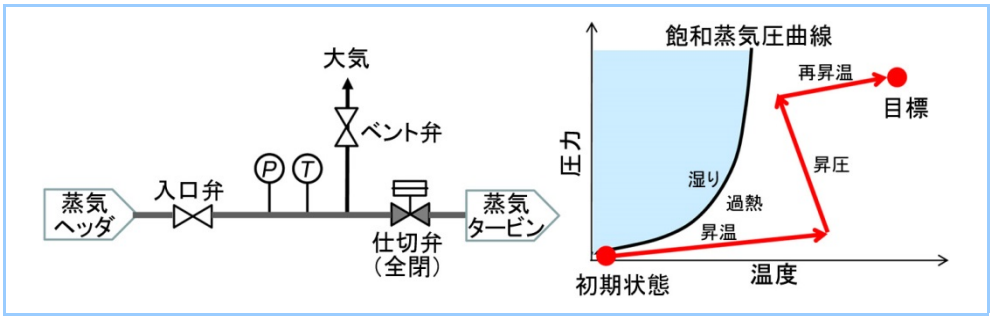


図3 対象操作の概要
学習対象とした蒸気配管暖気操作の構成と操作手順を示す

3.2 学習結果

大気温度が異なる夏季，冬季の2通りの運転条件で，配管内の蒸気状態に関する制約条件を満たす熟練者の基準操作データを設定し，これにベータ分布を与えてばらつきを考慮した操作データを生成し，教師データとした。

より広範な運転条件への適用を可能とするため，大気温度，及び教師データを夏季・冬季の条件交互に切替え，逆強化学習を行った。

冬季の運転条件へ適用したときの状態量(温度，圧力)と操作量(ベント弁開度，入口弁開度)の軌跡を図4に示す。配管内の蒸気の過熱状態を維持しながら温度，圧力を目標値まで昇温，昇圧できていること，並びにベント弁と入口弁の操作の軌跡が熟練者の段階的な操作に近い結果が得られており，熟練者の操作を習得できていることを確認した。

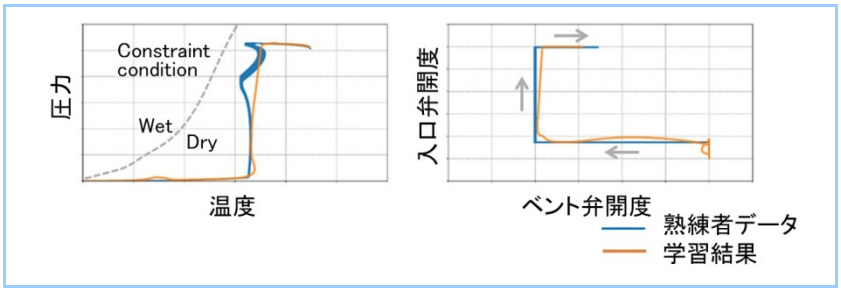


図4　逆強化学習結果と熟練者操作の運転軌跡比較
逆強化学習により熟練者操作を模倣した操作ができたことを示す

3.3　複数運転条件への適用

3.2 で得られた操作手順を夏季と冬季の大気温度の条件内挿する 10 ケースに適用したときの状態量・操作量の軌跡を図5に示す。全ケースについて制約条件を満足し，学習と異なる運転条件にも適用可能な操作手順となることを確認した。

同様に，夏季の条件，及びその教師データのみを用いて逆強化学習して得られる操作手順を上記の 10 ケースに適用したところ，大気温度が低い条件では目標値まで昇温・昇圧できなかった。

逆強化学習において，広範な運転条件に適用可能な操作手順を学習するには，異なる運転条件での熟練者データを用いた学習方法(同時学習)が有効であることを確認した。

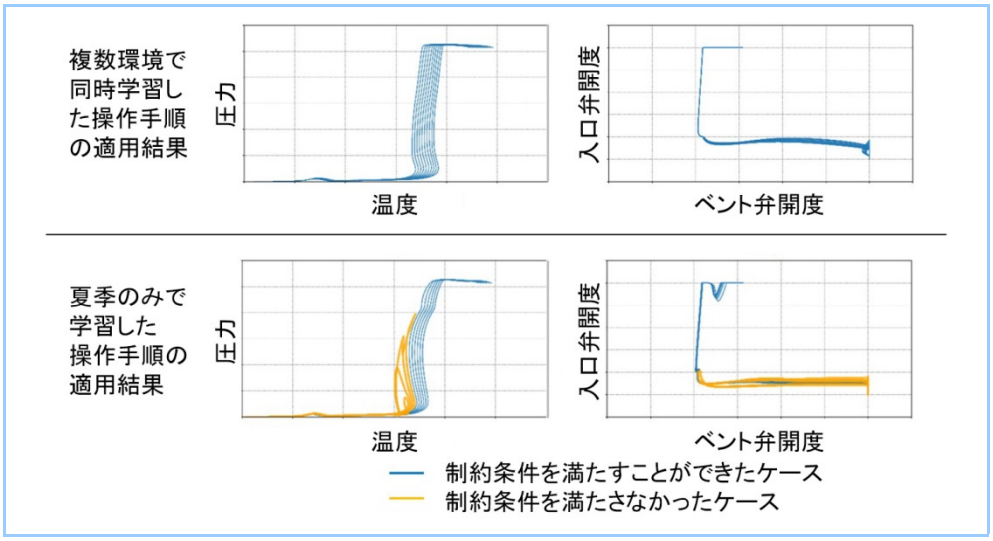


図5　学習した操作手順の複数運転条件への適用結果
学習時と異なる運転条件への適用可否を評価した結果を示す

3.4　報酬の可視化

逆強化学習で得られた評価指標(報酬関数・状態価値関数)について，ベント弁と入口弁の開度に対する評価値を可視化した。2つの弁開度以外の観測量である配管内の圧力や温度，流量等の状態量は2つの弁開度から決まる整定値を与えた。

夏季と冬季の同時学習で得られた関数の夏季，冬季の条件における可視化結果，並びに夏季・冬季の単一運転条件での学習で得られた関数のそれぞれの条件における可視化結果を図6に示す。

状態価値関数について，いずれのケースも入口弁が全開付近，すなわち昇圧完了し再昇温する操作終端付近の状態価値が高いことを学習している。一方，夏季の単一運転条件では，ベ

ント弁を閉めたのち，入口弁を全開にしていく昇圧過程における状態価値の変化が他のケースに比べて小さくなっている。昇圧操作の後半に行う入口弁の全開操作は，運転点が飽和蒸気圧曲線に近づき過熱度の余裕が小さくなる。過熱度の余裕を戻すため，昇圧完了後に再昇温への切替操作が重要となる。夏季は冬季に比べて温度が高く，過熱度の落ち込みが小さいため，この操作が冬季に比べて容易となり，3.3 節に示すように，夏季の単一運転条件での学習結果が冬季条件に適用できなかった要因の1つと考えられる。

報酬関数について，すべてのケースにおいて，熟練者による弁操作の軌跡にそった状態が他の状態に比べて相対的に高くなっている。特に，ベント弁が全開で，入口弁を全閉から徐々に開けていく操作は更に高い傾向を示しており，初期段階で過熱状態を作り出す操作の習得などに寄与したと考える。

状態価値関数と報酬関数を分けて学習できるAIRLは，学習結果の可視化に有効であり，熟練者のノウハウの推量に繋がることを確認した。

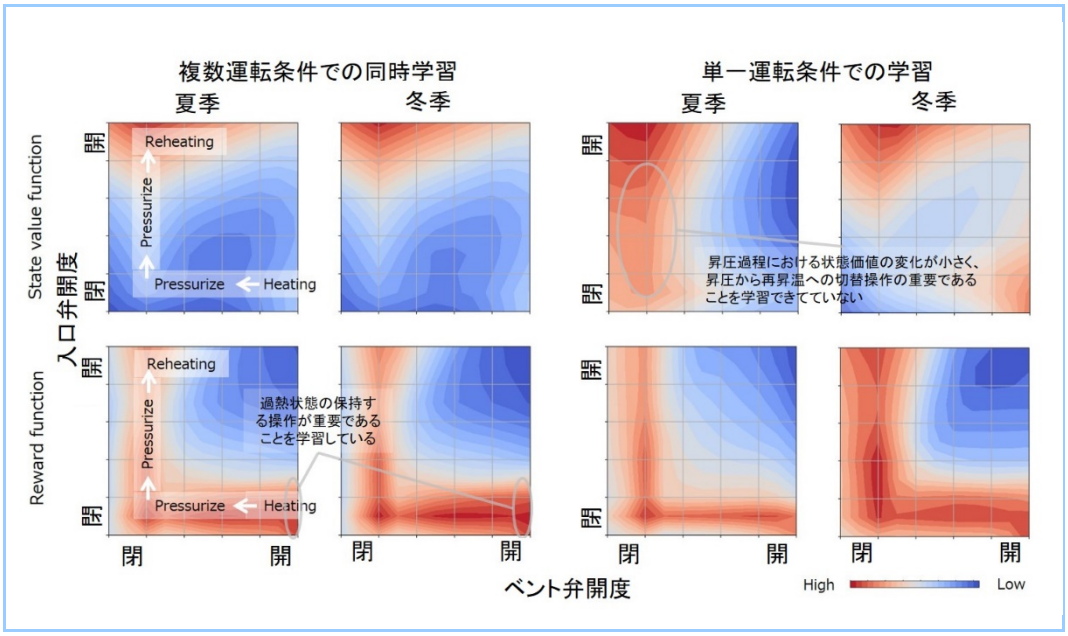


図6 評価指標の可視化
弁開度を変化させた際の評価指標出力値のマップを示す

4. まとめ

本報では，複数運転条件の熟練者の操作データを交互に(同時に)与えて逆強化学習することで，広範な運転条件に適用可能で熟練者の操作手順を習得できること，熟練者のノウハウ推量に繋がる評価指標が得られることを紹介した。

本技術と当社が有するプラントシミュレーション技術との組合せにより，定量的な運転操作指針を生成でき，手動操作を要するプラント・機器の運転支援や自動化への展開が可能となる。

今後は，実機で求められるより複雑な問題に対応するため，計器情報以外にカメラ画像や運転員の視線情報等を用いたマルチモーダルな情報を用いた逆強化学習技術の開発を計画している。

参考文献

(1) D. Silver, et al., A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play, Science. Vol.362, No.6419, pp.1140-1144, 2018.
(2) H. Yoo, et al., Reinforcement learning based optimal control of batch processes using Monte-Carlo deep deterministic policy gradient with phase segmentation, Computers and Chemical Engineering. 315 Vol.144, 107133, 2021.
(3) J. Fu, et al., Learning Robust Rewards with Adversarial Inverse Reinforcement Learning, arXiv. org. arXiv:1710. 11248, 2018.