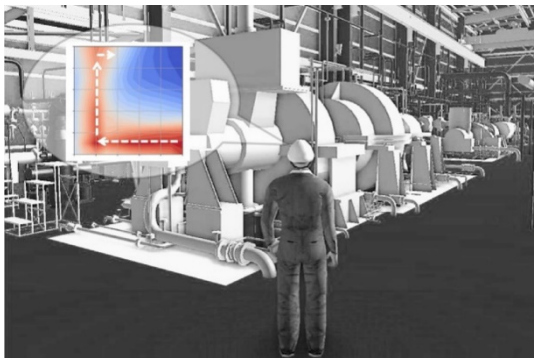
Inverse Reinforcement Learning Technology Contributing to Imitation and Know-How Visualization of Plant Expert Operations



YOSUKE NAKAGAWA[*1] HITOI ONO[*2]

YUSUKE HAZUI[*1] SACHIYO ARAI[*3]

Non-routine manual operations of a plant require appropriate judgment according to the operating state, and significantly depend on the knowledge and skill of the operator. Deep reinforcement learning can recognize a state and learn a series of operations based on an evaluation index for that state. It is expected as a method that leads to the clarification of operational guidelines, which have been considered tacit knowledge thus far. However, due to the problems such as the difficulty in designing the evaluation indexes required for learning, the explainability of the learning results, etc., its application to plant operations, which requires safety and reliability, has not progressed. This report presents the acquisition of expert operations using inverse reinforcement learning technology developed in collaboration with Chiba University and the technology to surmise the know-how of experts through visualization of the learning results.

1. Introduction

Non-routine manual operations such as starting and stopping of a plant require changing the timing and amount of operation according to the plant state, which significantly depends on the knowledge and skill of the operator. In the past, in order it was common to describe the operation procedures of experts in the if-then rule form depending on each plant state and establish an operating system to assist automation and inexperienced operators to reduce the influence of the operator on the plant. However, it is not easy to describe the operation procedures of experts in a rule form because such operations often contain tacit knowledge.

In recent years, with the improvement of computer speed, deep learning for situational awareness, has been developed. Deep reinforcement learning, a combination of this deep learning and reinforcement learning, which is responsible for acquiring operations (action) in response to the situation, is expected as a method of acquiring appropriate action by maximizing the evaluation index (reward) of a series of operations, for example, beginning to completion of startup. In the field of competitive games, AlphaZero[(1)], a leading computer program, has proved that deep reinforcement learning can achieve performance superior to that of humans, and it is expected that deep reinforcement learning be also applied [(2)] to the field of plant operation.

On the other hand, there are problems that it is not easy to design a reward for acquiring the operation of experts and that the obtained learning result needs to be explainable, as well as problems with acquisition of robust operating guidelines that can be applied to a wider range of operating conditions.

In order to solve these problems, we developed inverse reinforcement learning technology that can estimate the reward from the training data of expert operations and acquire the operation of the experts, and conducted verification through the simulation using a warming operation of piping with steam as an example. As a result, we were able to acquire the operation procedure of experts, and surmised the know-how of experts by visualizing the obtained reward. This report presents our efforts.

*1 CIS Department, Mitsubishi Heavy Industries, Ltd.
*2 Chief Staff Manager, CIS Department, Mitsubishi Heavy Industries, Ltd.
*3 Professor, Chiba University

Thereafter, chapter 2 provides an overview of inverse reinforcement learning, chapter 3 describes the learning and visualization results from the steam pipe warming operation, and chapter 4 provides a conclusion.

2. Inverse reinforcement learning technology

2.1 Reinforcement learning and inverse reinforcement learning

Reinforcement learning, which is one of the machine learning methods, is to learn operation procedures that maximize the reward (evaluation index) obtained from the environment through trial and error, maximizing not the reward for each operation, but the cumulative reward in a series of operations. Applying reinforcement learning to plant operations entails problems such as that the reward design for a problem to be solved is not easy, that new operation procedures that cannot be expected by plant designers and operators may have new risks that are not found by the calculation, and that the explainability for the obtained learning results and operation procedures is required.

Inverse reinforcement learning is a technology that enables estimation of rewards as well as acquisition of operation procedures with use of the operating data of experts to deal with these problems. There have long been technologies that simply imitate the operation of experts, such as imitation learning. On the other hand, inverse reinforcement learning can undertake learning even under operating conditions where no reward is designed and there is no operation data of experts, by transferring the estimated reward to operating conditions different from those at the time of learning, and performing new reinforcement learning using it. As a result, it may be possible to learn operation procedures that can be applied to a wide range of operating conditions with a small amount of data and the number of trials.

2.2 Developed inverse reinforcement learning technology

In the efforts presented in this report, we used adversarial inverse reinforcement learning (AIRL)[(3)], a method of inverse reinforcement learning incorporating a generative adversarial network, which is one of the deep learning technologies. It is said that this method can learn more complicated action compared to conventional inverse reinforcement learning. **Figure 1** shows the configuration of AIRL.

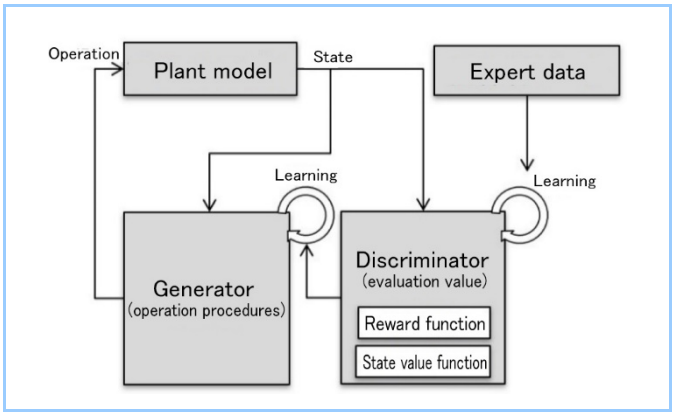


Figure 1 Configuration of AIRL
Configuration of learner of AIRL, adversarial inverse reinforcement learning

The adversarial inverse reinforcement learning consists of two neural networks, i.e., generator and discriminator. The generator learns the operation procedures by reinforcement learning, inputs the procedures to the plant, and generates operation data (state transition history). The discriminator discriminates between the generated-by-learning operation data (false data) and the expert operation data (true data). The generator learns operation procedures that imitate experts more in an attempt to deceive the discriminator, while the discriminator learns to discriminate the authenticity more so as not to be deceived by the generator. When the generator has learned expert operation, the discriminator cannot discriminate between the generator's operation and experts' operation. This means that the generator has completed the learning to imitate the action of experts.

The AIRL discriminator learns the evaluation index of operation procedures by dividing it into two functions: reward function and state value function. The reward function represents the

relative goodness of the operation in each state. The state value function represents the cumulative reward from the initial state and indicates the final goal. By visualizing these functions, the know-how of experts can be surmised, which leads to explanation of the learning results.

On the other hand, the operation procedures that can be obtained by inverse reinforcement learning are limited to an operation neighborhood of expert given as training data. Applying the operation procedures to a wider range of operating conditions requires re-learning under the new conditions. In the work presented in this report, we extended the functionality of AIRL and developed a simultaneous learning method to learn multiple operating conditions alternately (simultaneously). **Figure 2** shows the configuration diagram.

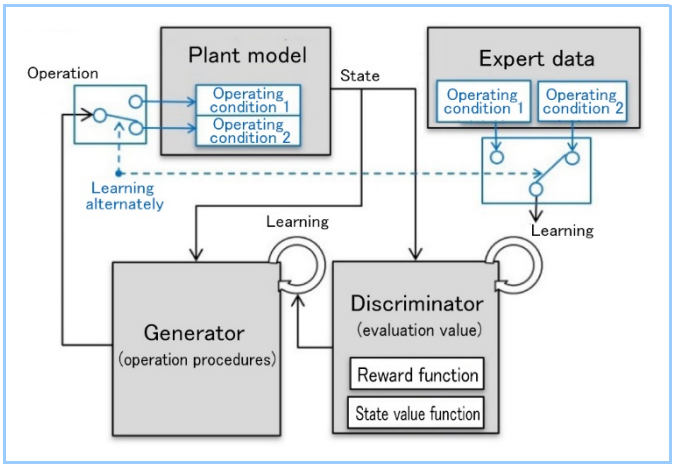


Figure 2 Configuration of developed inverse reinforcement learner
The conventional AIRL function was extended to learn multiple operating conditions alternately.

3. Verification of developed technology with plant simulator

3.1 Overview of target operation

Figure 3 shows the steam pipe warming operation. This operation uses vent valves and inlet valves to increase the temperature and pressure in the piping from atmospheric conditions to a state close to the source pressure and temperature with a smaller consumption amount of high-temperature and high-pressure steam supplied from the upstream plant. After the warming is completed, the steam is supplied to turbines and other equipment. For this operation, the steam in the piping must be kept superheated so that it does not contain droplets such as mist in order to prevent problems such as turbine blade damage and hammering in the piping. Therefore, experts manually perform the operation in steps: (1) slightly opening the inlet valve with the vent valve fully open (to rise the temperature), (2) closing the vent valve to a slight opening and then fully opening the inlet valve (to rise the pressure), and (3) opening the vent valve to an intermediate opening with the inlet valve kept fully open (to re-rise the temperature).

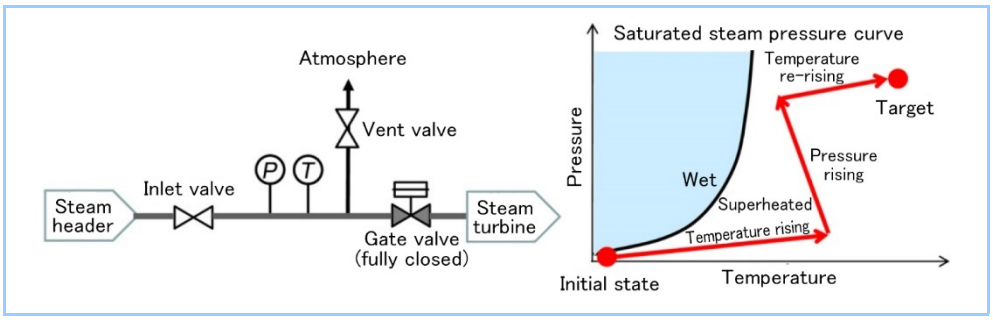


Figure 3 Outline of target operation
Configuration and procedures of steam pipe warming operation, which is learning target

In addition, experts perform the temperature and pressure rising operations with less steam consumption by changing the valve operation amount and operation timing in consideration of the difference in heat loss amount due to the atmospheric temperature.

3.2 Learning result

We set up the reference expert operation data that satisfied the constraint conditions regarding the steam state in the pipe under two different operating conditions with different atmospheric temperatures (summer and winter), gave beta distribution to the data to generate the operating data considering the variation, and used these as training data.

In order to enable the operation procedures to be applied to a wider range of operating conditions, we performed inverse reinforcement learning while alternating the summer and winter conditions of the atmospheric temperature and training data.

Figure 4 shows the trajectories of the state quantity (temperature and pressure) and operation amount (vent valve opening and inlet valve opening) when the operation is applied to winter operating conditions. The temperature and pressure could be risen to the target values while maintaining the superheated state of steam in the piping, and the obtained trajectories of the operation of vent valve and inlet valve were close to the stepwise operation of experts. In this way, it was confirmed that the operation ability of experts could be acquired.

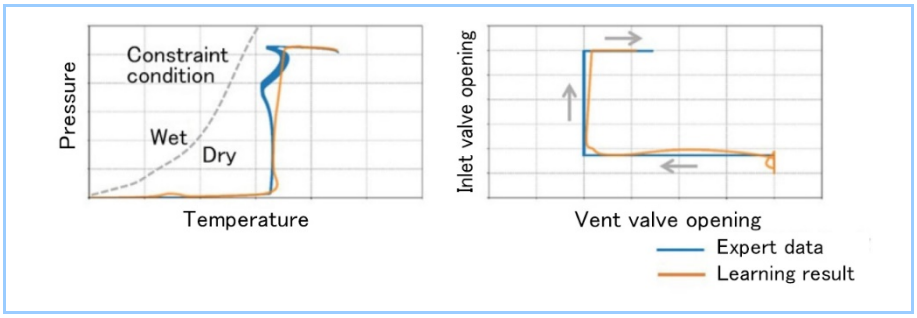


Figure 4 Comparison of trajectories between inverse reinforcement learning result and expert operation

It was confirmed that the inverse reinforcement learning could imitate the expert operation.

3.3 Application to multiple operating conditions

Figure 5 shows the trajectory of the state quantity and operation amount when the operation procedures obtained in 3.2 are applied to 10 cases in which the atmospheric temperature conditions in summer and winter are interpolated. The constraint conditions were satisfied in all cases, so it was confirmed that the operation procedures can be applied to operating conditions different from those in learning.

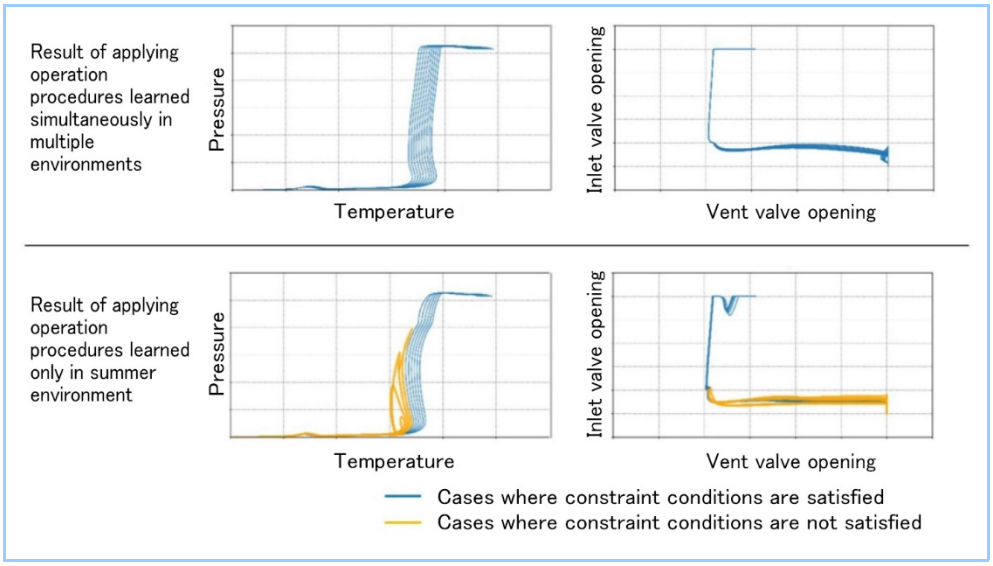


Figure 5 Results of applying learned operation procedures to multiple operating conditions

Result of evaluating applicability to operating conditions different from those in learning

On the other hand, when the operation procedures obtained by inverse reinforcement learning using only the summer conditions and training data were applied to the above 10 cases, it was not possible to rise the pressure and temperature to the target value under low atmospheric temperature

conditions.

It was confirmed that a learning method using expert data under different operating conditions (simultaneous learning) is effective for acquisition of operation procedures that can be applied to a wide range of operating conditions from inverse reinforcement learning.

3.4 Visualization of reward

We visualized the evaluation values with respect to the openings of the vent valve and the inlet valve resulted from the evaluation index (reward function and state value function) obtained by inverse reinforcement learning. The observed quantities other than the two valve openings, such as in-pipe pressure, temperature, and flow rate, were given set values determined by the two valve openings.

Figure 6 shows the visualization results under the summer and winter conditions from the function obtained from the simultaneous learning of summer and winter conditions, as well as the visualization results under each of the summer and winter conditions from the function obtained from the single learning of that operating condition.

The state value function has learned that the state value is high when the inlet valve is nearly full-open, that is, when the operation is near the end where rising the pressure has been completed and the temperature is re-risen in each case. On the other hand, under summer operating condition alone, the change in state value during the pressure rising process in which the vent valve is closed and then the inlet valve is fully opened is smaller than in other cases. The operating point in the fully open operation of the inlet valve performed in the latter half of the pressure rising operation approaches the saturated vapor pressure curve and the margin of superheat degree becomes small. In order to recover the margin of superheat, it is important to switch to re-rising the temperature after increasing the pressure is completed. In summer, this operation is easier than in winter because the temperature is higher and the drop in superheat is smaller than that in winter, which is considered to be one of the reasons why the learning results with the operating condition in summer alone could not be applied to winter conditions, as shown in section 3.3.

Regarding the reward function, the evaluation values in the state in line with the trajectory of expert valve operation were relatively higher than those in other states in all cases. In particular, the operation, in which the inlet valve is gradually opened from fully closed with the vent valve fully opened, shows a tendency for the evaluation value to be even higher and is considered to be the contribution to the acquisition of the operation to create superheated conditions in the initial stage.

It was confirmed that AIRL, which can learn the state value function and the reward function separately, is effective for the visualization of learning results and leads to the estimation of expert know-how.

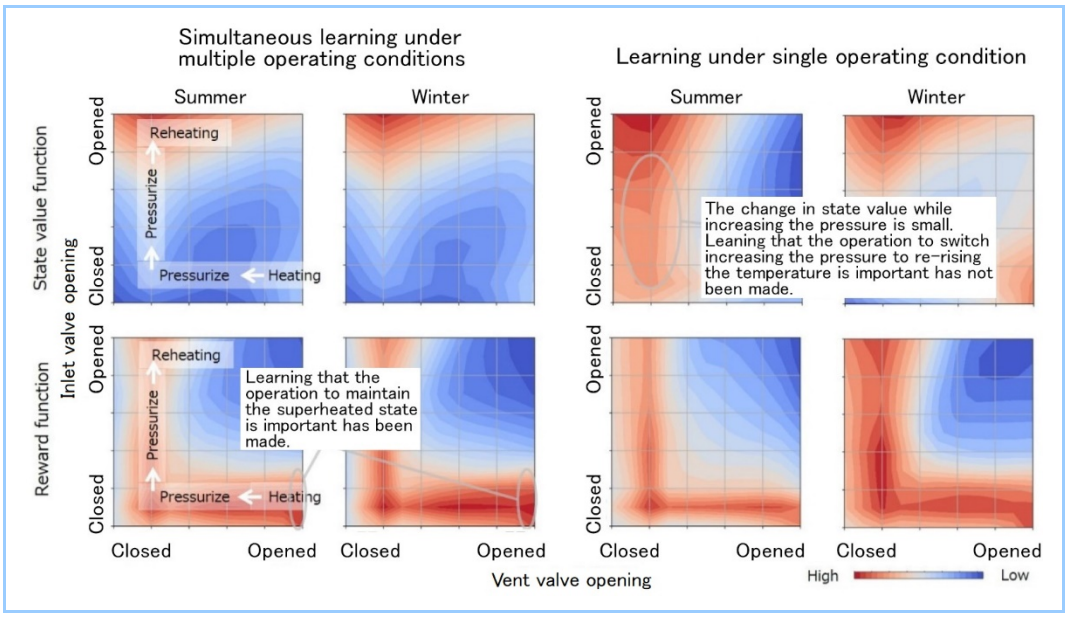


Figure 6 Visualization of evaluation indexes
Maps of evaluation index output values with regard to changing valve opening

4. Conclusion

This report explained that it is possible to acquire expert operation procedures that can be applied to a wide range of operating conditions, and to obtain evaluation indexes that lead to the estimation of know-how of experts by performing inverse reinforcement learning while giving expert operation data of multiple operating conditions alternately (simultaneously).

Combining this technology with our plant simulation technology makes it possible to generate quantitative operation guidelines and to develop them to operation support and automation of plants and equipment that require manual operation.

Moving forward, we plan to develop inverse reinforcement learning technology that uses multimodal information such as camera images and operator's line-of-sight information in addition to instrument information in order to deal with more complicated requirements of actual machines.

References

(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.