Preliminary study on evacuation infrastructure optimization through reinforcement learning using policy-based algorithms

Joonseok Lim[a], Minho Hwang[a], Geon Kim[a], Sungmin Han[a], Gyunyoung Heo[a]
[a]Department of Nuclear Engineering, Kyung Hee University, Yongin-si, Republic of Korea

Abstract: Residents should evacuate urgently in the event of a disaster, including a radiation emergency. Evacuation path optimization in emergency situations is a difficult problem to solve due to dynamic, and complex constraints, and large uncertain phenomena, but optimization is being attempted through reinforcement learning in many studies. On the other hand, the role of infrastructure mitigating the consequence of emergency situation is clearly significant, while their optimization in terms of operational viewpoint was not easily recognized. The platform for radiological emergency agent-based integrated simulation model (PRISM) is an agent-based model for simulation of wide-area evacuation during radiation emergency situations. PRISM includes a resident evacuation algorithm using a path finding model, atmospheric diffusion of radioactive materials, and interaction between infrastructure and evacuees, and is being updated for realistic simulation. The purpose of this paper is to suggest the method how to optimize the radiation emergency response strategy, that is, the infrastructure operational strategy, through this platform to achieve increasing the recovery level (REC) value. It was shown in the previously mentioned evacuation path optimization that reinforcement learning can be a key in this situation. Previous studies have applied to one infrastructure with a value-based algorithm, Deep Q-Network (DQN). In this study, a policy-based reinforcement learning algorithm was applied to solve the shortcomings of DQN. Results suggest that when an evacuation simulation is performed by applying the distribution of infrastructures optimized through reinforcement learning, the REC value increases faster compared to an evacuation simulation performed with the default (uniform) distribution.

Keywords: Radiological evacuation, Infrastructure operation, Reinforcement learning

1. INTRODUCTION

Emergency evacuation is a difficult problem to optimize due to highly dynamic variables and complex constraints [1]. Matters to be considered for optimizing emergency evacuation include the location and number of destinations, the shortest distance to the destination, and the presence or absence of obstacles on the route [1-4]. In these studies, reinforcement learning was used to optimize emergency evacuation. This is because reinforcement learning has strengths in optimizing path planning at a large scale that encompasses the preceding considerations [5].

A nuclear or radiological accident is one of those accidents that can cause long-term, widespread and serious impacts, requiring emergency evacuation. General safety requirements for radiological emergency response preparedness (EPR) are provided in IAEA GSR Part 7 [6]. One of the important elements covered in this document is the capability of infrastructure. Therefore, in this paper, the infrastructure is optimized through reinforcement learning, rather than path planning, which was addressed in previous studies in the emergency evacuation field.

2. REINFORCEMENT LEARNING

A reinforcement learning problem can be expressed as a system consisting of an agent and an environment. The environment creates information that represents the state of the system. Agents interact with the environment by selecting actions using information obtained by observing the state. Through the agent's actions, the environment transitions to the next state, and at this time, a defined reward is paid to the agent. The cycle of “State>Action>Reward” means that one time step has passed, and this cycle is repeated until the end of the environment or until a specific state defined by the user is reached. The figure 1 showing this is as follows [7].

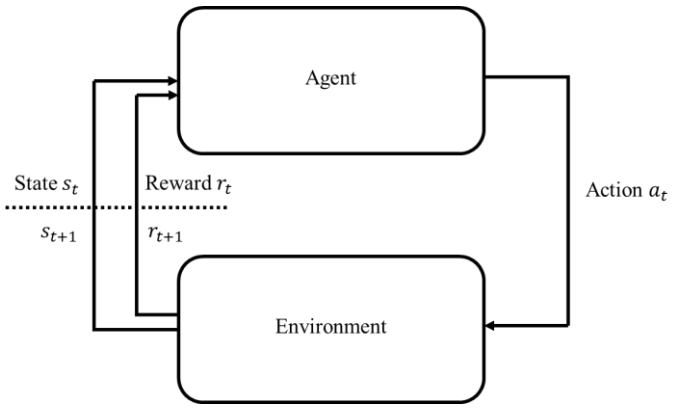

Figure 1. The cycle of reinforcement learning

A policy-based algorithm is an algorithm that learns a policy (π), which is defined as a function that maps states to action probabilities. As the value-based algorithm, the goal of this algorithm is to maximize the expected value of the cumulative discounted reward, but the difference is that it updates the policy, in contrast to the value-based algorithm that updates the value. The policy gradient, $\nabla_\theta J(\pi_\theta)$ is used to update the policy, which is expressed as equation 1, where θ refers to the neural network. τ is defined as a trajectory, which means cycle $((a_0, s_0, r_0), \ldots, (a_T, s_T, r_T))$ as shown in figure 1 until the end time (T) of one episode (or simulation). The larger the return ($R_t(\tau)$) and the greater the probability of action (a_t) for a state (s_t), the higher the policy gradient. Ultimately, the goal of a policy-based algorithm is to find the maximum value of $J(\pi_\theta)$ where the policy gradient becomes 0. There are two major advantages of policy-based algorithms: 1) it can simulate the continuous action space, and 2) it can represent stochastic policies.

$$\nabla_\theta J(\pi_\theta) = E_{\tau\sim\pi_\theta}[\sum_{t=0}^{T} R_t(\tau)\nabla_\theta log\pi_\theta(a_t|s_t)] \quad (1)$$

3. ALGORITHM SETTING

3.1. Environment

Platform for Radiological emergency agent-based Integrated Simulation Model (PRISM) is an agent-based platform for simulation of wide-area evacuation during radiation emergency situations [8]. PRISM includes a resident evacuation algorithm using a path finding model, atmospheric diffusion of radioactive materials derived from HYSPLIT, infrastructure-infrastructure interaction, and infrastructure-evacuee interaction, and is being updated for realistic simulation. The purpose of this paper is to optimize the radiation emergency response strategy, that is, the infrastructure commitment strategy, through this platform to achieve the fastest evacuation time, thereby quickly increasing the recovery level (REC) value, which indicates resilience. REC is hypothesized using a function $g(\cdot)$ as equation 2. The added infrastructure is police (i.e., traffic control capability) and has the effect of solving traffic jams and speeding up evacuees. This shortens the evacuation time and quickly increases REC.

$$REC(t) = g(S(t), H(t), I(t)) \quad (2)$$

where, $S(t)$ refers to the amount representing the damage resistance or recovery ability of the recovery target, $H(t)$ refers to the amount representing the damage caused by the hazard element, and $I(t)$ refers to the amount related to the mitigation infrastructure element.

3.2. Components of Reinforcement Learning

The area of interest is set to four roads as shown in the figure 2, and a set of the number of evacuees on ith road at time t, n_t^i is defined by the state, S_t as shown in the equation 3. At this time, the number of evacuees does not include evacuees who have completed evacuation.

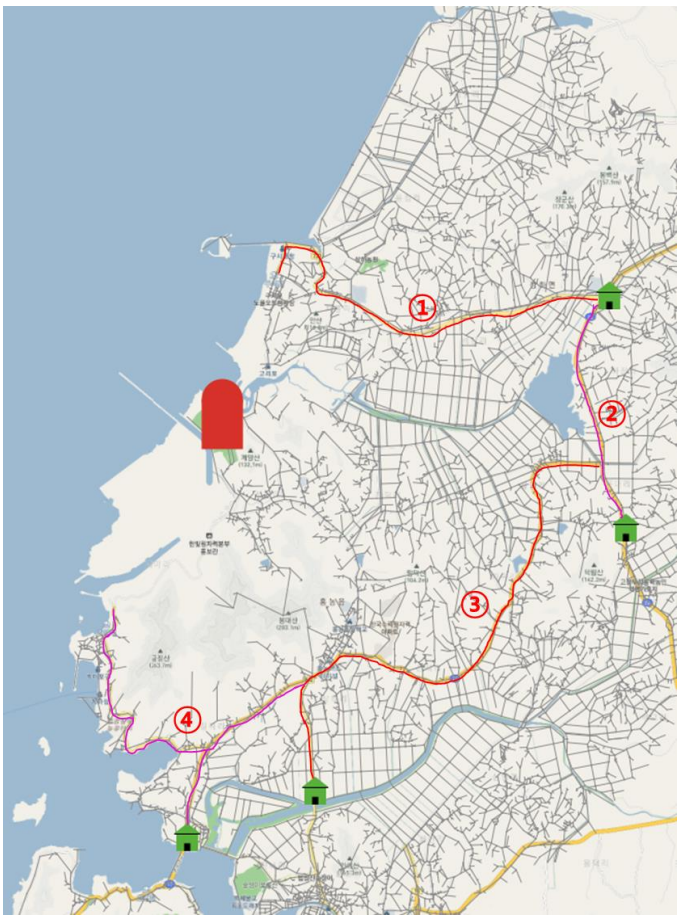

Figure 2. The overview of the environment

$$S_t = [n_t^1, n_t^2, n_t^3, n_t^4] \tag{3}$$

A neural network that receives the state as input derives the number of polices on each road as an action. Initially, the assumed number of polices is evenly distributed. Reward $r_{t'}$ is defined as the number of evacuees who completed evacuation from $t-1$ to t. There are two constraints. First, the total number of polices is constant. Second, if the number of police is derived as a negative number, it is replaced by 0. These constraints operate as negative rewards $r_{t'}^c$, and a reward equal to -1% of the initially set total number of evacuees is assigned to the agent. The return $R_t(\tau)$, which represents the sum of rewards to which the discount factor γ is applied, is shown in equation 4.

$$R_t(\tau) = \sum_{t'=t}^{T} \gamma^{t'-t}(r_{t'} + r_{t'}^c) \tag{4}$$

4. RESULTS

The number of evacuees was assumed to be 1000 and the total number of polices was assumed to be 200. The number of evacuees is related to population density, but since the focus of this paper is to determine the applicability of reinforcement learning, population density was not considered. It was also assumed in the same context as the number of evacuees in the number of polices. If the number of police is small, traffic jams will not be resolved, and the time required to complete evacuation will increase. Conversely, it is expected that the time required to complete evacuation will decrease and then converge at a certain level.

In figure 2, the place represented by the green house is the final destination, that is, a shelter, and the place shown in red rounded rectangle represents the nuclear power plant. Evacuees are randomly generated on the road on the map and each take the shortest distance to their destination. Learning progressed through 130 iterations, and the total reward the agent received as learning progressed is shown in the figure 3. The figure 4

is a graph comparing the average speed of evacuees when police were deployed through the learned neural network and when police were not deployed.

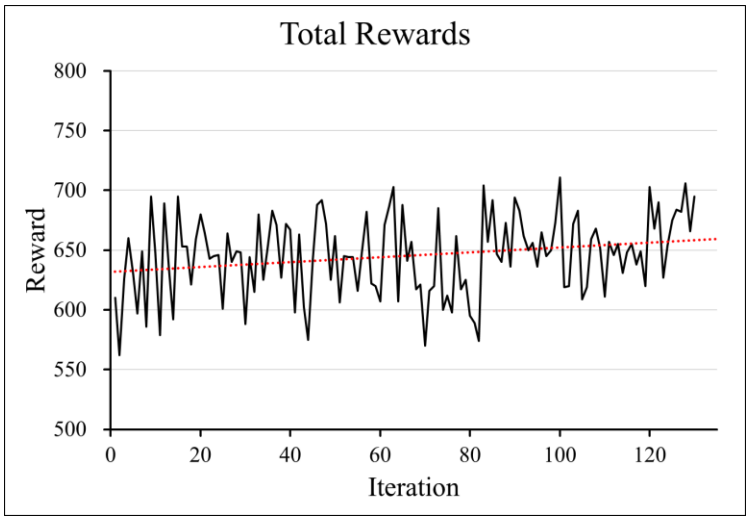

Figure 3. The total reward of REINFORCEMENT algorithm

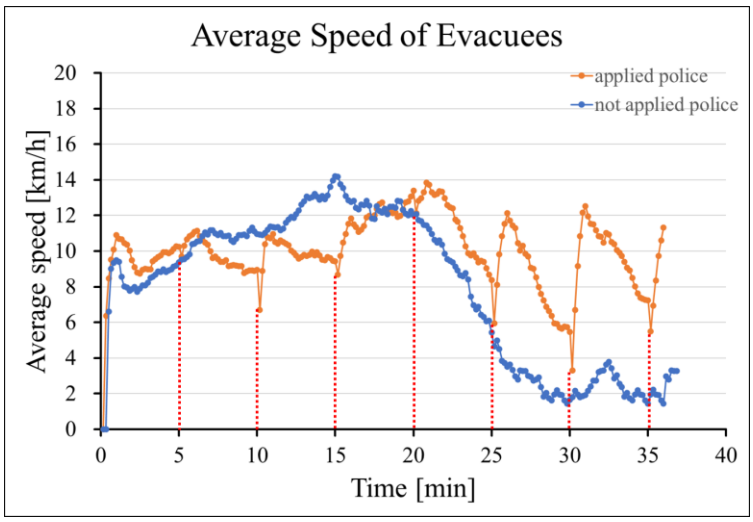

Figure 4. The comparison of average speeds of evacuees. The red dash lines mean the time to change the police infrastructure.

Total rewards appear to increase as learning progresses, but the increase is not large in figure 3. This is due to the sample-inefficient characteristic of the training process, which is one of the disadvantages of REINFORCEMENT algorithm [9]. This characteristic requires a large number of samples to obtain an information. According to Figure 4, it can be seen that the average speed of evacuees increases for each unit of time (5 min) that the police are deployed. However, there appeared to be no significant difference in evacuation completion time.

5. CONCLUSION

The infrastructure targeted for optimization in this study was the traffic control using a notion of ‘police,’ which affects the evacuation speed of evacuees. Since there are countless types of police distribution per unit of time, finding the optimal distribution is a difficult problem. An attempt was made to optimize the distribution of polices using reinforcement learning, which is mainly used for optimization problems in emergency evacuation situations. Assuming four roads, the number of evacuees on each road was set to state, the number of polices was set to action, and the number of evacuations completed and constraints were set to reward. In previous research, there was a case in which infrastructure distribution was optimized using DQN, one of the value-based algorithms [10]. However, the disadvantages of value-based algorithms are that they only simulate a discrete action space and that a stochastic policy cannot be derived.

Optimization methods using neural networks, including reinforcement learning using policy-based algorithms, have the characteristics of a black box in which it is unclear what process produces the output. Additionally, it cannot be guaranteed that the solution obtained through the reinforcement learning algorithm used in this paper is the global minimum unless it goes through a significant number of iterations. However, in a radiation emergency with a large sequence, a little help for evacuation can play a big role. In future research, the number of infrastructures will be increased to derive the optimal distribution of each infrastructure. Additionally, the applicability of other policy-based algorithms will be investigated.

Acknowledgements

This work was supported by the Nuclear Safety Research Program through the Regulatory Research Management Agency for SMRS (RMAS) and the Nuclear Safety and Security Commission (NSSC) of the Republic of Korea (No. 1500-1501-409) and the Human Resources Development of the Korea Institute of Energy Technology Evaluation and Planning(KETEP) grant funded by the Korea government Ministry of Knowledge Economy (No. RS-2023-00244330).

References

[1] Sharma, J., Andersen, P. A., Granmo, O. C., & Goodwin, M. (2020). Deep Q-learning with Q-matrix transfer learning for novel fire evacuation environment. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 51(12), 7363-7381.

[2] Xu, D., Huang, X., Mango, J., Li, X., & Li, Z. (2021). Simulating multi-exit evacuation using deep reinforcement learning. Transactions in GIS, 25(3), 1542-1564.

[3] Zhang, Y., Chai, Z., & Lykotrafitis, G. (2021). Deep reinforcement learning with a particle dynamics environment applied to emergency evacuation of a room with obstacles. Physica A: Statistical Mechanics and its Applications, 571, 125845.

[4] Li, X., Liu, H., Li, J., & Li, Y. (2021). Deep deterministic policy gradient algorithm for crowd-evacuation path planning. Computers & Industrial Engineering, 161, 107621.

[5] Zheng, S., & Liu, H. (2019). Improved multi-agent deep deterministic policy gradient for path planning-based crowd simulation. IEEE Access, 7, 147755-147770.

[6] International Atomic Energy Agency. (2015). PREPAREDNESS AND RESPONSE FOR A NUCLEAR OR RADIOLOGICAL EMERGENCY. IAEA SAFETY STANDARDS SERIES No. GSR Part 7. IAEA. Vienna.

[7] Graesser, L., & Keng, W. L. (2019). Foundations of deep reinforcement learning: theory and practice in Python. Addison-Wesley Professional.

[8] Kim, G., & Heo, G. (2023). Agent-based radiological emergency evacuation simulation modeling considering mitigation infrastructures. *Reliability Engineering & System Safety*, *233*, 109098.

[9] Sutton, R. S., McAllester, D., Singh, S., & Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. Advances in neural information processing systems, 12.

[10] Kim, G. B., & Heo, G. Y. (2022). A study on finding an optimal response strategy considering infrastructures in an agent-based radiological emergency model using a deep Q-network. ESREL2022.