Hierarchical Reinforcement Learning for Epidemics Intervention

Yinzhao Dong
Dalian University of Technology
Dalian, China
1447866357@qq.com

Chao Yu
Sun Yat-sen University
Guangzhou, China
yuchao3@mail.sysu.edu.cn

Lijun Xia
Sun Yat-sen University
Guangzhou, China
xialj5@mail2.sysu.edu.cn

ABSTRACT

Epidemics of infectious diseases are an important threat to public health and global economies. The intervention in epidemics is essential, which aims to minimize the total number of infected people and, at the same time, minimize the amount of intervention on human mobility. However, the complexity of pre-symptomatic patients and dynamic human mobility cause significant challenges for developing efficient intervention strategies for epidemic diseases. For this reason, we first use the visit history of each area to calculate the health probability of each person, so as to find the most likely pre-symptomatic patients. Then, a Hierarchical Reinforcement Learning Intervention (HRLI) model is proposed to automatically learn intervention strategies based on the information gathered from the individuals and the area of residences. The model can take the advantage of hierarchical supervision to realize the macro-control in an area of residence and precise intervention for each individual in this residence, respectively.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

Epidemics Intervention, Hierarchical Reinforcement Learning, Sparse Reward

ACM Reference Format:
Yinzhao Dong, Chao Yu, and Lijun Xia. 2020. Hierarchical Reinforcement Learning for Epidemics Intervention. In *KDD'20 WORKSHOP: ACM SIGKDD Conference on Knowledge Discovery and Data Mining, August 24, 2020, PAPW, California*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

PAPW, August 24, 2020, KDD'20 WORKSHOP, California

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The new corona-virus disease 2019 (COVID-19), a newly emerged epidemics disease, is a serious threat to public health and global economy. The long incubation period of this disease makes it difficult to identify the infected individuals, and leads to a large consumption of human resources and medical resources. Timely intervention is crucial for combating epidemics like COVID-19, in order to minimize the total number of infected people, at the same time, reduce the impact of interventions on people's daily life to the minimum. Based on this, we first use the visit history of each area to calculate the health probability of each person, so as to find the potential infected people. Then, a Hierarchical Reinforcement Learning Intervention (HRLI) model is proposed to automatically learn intervention strategies based on the information gathered from the individuals and the area of a residence. Specifically, the residence and each individual are model as two agents and their decision making processes as two *Markov Decision Processes* (MDPs). The two agents can coordinate the decision making processes through the *risk level*, which is define based on the infected number of the residence. The residence uses the DQN [2] algorithm to dynamically adjust the *risk level*. Meanwhile, the individual takes *risk level* of its of its residence as a dimension of state and trains with the PPO [3] to obtain the probability boundary of intervention actions. Our model has the following advantages:

- It can ultimately take the advantage of hierarchical supervision to realize the macro-control in an area of residence and precise intervention for each individuals in this residence.
- We use the *effective reproduction ratio* [1] to define the reward in each day so as to solve the problem of sparse reward in the epidemics.

in order to find an optimal intervention strategy.

2 THE HRLI MODEL OF EPIDEMICS

In this section, we first introduce the basic epidemiological simulation model. Then, we propose a health probability model, which uses the individual's visit history to calculate its health probability. Next, the MDPs for decision making in a residence area and each individual are introduced, respectively. Finally, we propose the HRLI model, in order to learn an effective epidemic intervention strategies.

2.1 The Epidemiological Model

The epidemiological model [4] aims to seek for the best intervention strategy based on five simulated scenarios of 10K-people over 60 days. The basic scenario assumes that each

individual goes to a working area during weekdays and visits a commercial area during weekends. Virus is transmitted with a certain probability when two people are in the same area. Except the basic scenario, there are another four scenarios where some changes have been made, such as: higher infection rates, larger number of areas, larger infected population at the beginning and longer working hours. More details of the simulation model can be found at the website of the Challenge on Mobility Intervention for Epidemics [5].

An individuals health status follows the stages as below: susceptible, pre-symptomatic, symptomatic, critical and recovered. Meanwhile, there are five intervention actions: no intervene, confine (in an area), quarantine (in the home), isolate (no contact with anyone), and hospitalize. The goal of the epidemiological model is to minimize the total number of infected people and the amount of intervention on human mobility. Therefore, the evaluation metric *Score* is a combination of the total number of interventions and infected people, as follows

$$Score = e^{\frac{I}{500}} + e^{\frac{Q}{10000}} \tag{1}$$

where I denotes the accumulated number of infected people on 60 days, $Q = N_{hos} + 0.5N_{iso} + 0.3N_{qua} + 0.2N_{con}$ denotes the weighted sum of the number of people being intervened, and N_{hos}, N_{iso}, N_{qua} and N_{con} denote the number of hospitalized, isolated, quarantined and confined people, respectively.

2.2 The Healthy Probability Model

The difficulty of epidemiological model mainly lies in how to identify the pre-symptomatic infected patients and which intervention to take for the undiscovered (including pre-symptomatic and symptomatic) people. Therefore, an infection probability model is proposed to calculate the current health probability P of each person. The specific process is as follows:

Step 1: Determine the current health status and infection status of the 10K-individuals at the start of one day. The health probability $\hat{p}$ of all individuals is initialized to 1.

Step 2: Obtain the current discovered (suspected or infected) set and healthy (susceptible or pre-symptomatic) set of people.

Step 3: The health probability of all individuals in the discovered set is set to 0.

Step 4: In order to cut off the spread of the epidemic as quickly as possible, all individuals in the discovered set are send to be hospitalized or isolated. Only those who become recovered can regain freedom.

Step 5: Obtain the visit history in the past 5 days for each area. Then, pick out the set of discovered individuals Set_{dis} and the set of healthy individuals Set_{hea} in each hour, respectively.

Step 6: The health probability $\hat{p}_i$ of each individual i in the Set_{hea} can be expressed as follows:

$$\hat{p}_i = \hat{p}_i * (1 - \frac{p_s * N_{dis}}{N_{hour} + e^{-7}}) \tag{2}$$

where p_s denotes the probability that an individual can be infected from a stranger contact, N_{dis} and N_{hour} denote the number of discovered individuals and healthy individuals, respectively.

Step 7: For each infected individual, the Set_{hea} can be divided into acquaintances and strangers. The probability p_c being infected by acquaintances is much higher than p_s. Using the Eq (2), the health probability of acquaintances will be higher than the actual value, which would increase the spread of the epidemic. In order to make this model more accurate, the health probability of acquaintances should be adjusted as follows:

$$\hat{p}_j = \hat{p}_j * (1 - p_c) \tag{3}$$

where j denotes an acquaintance of the Set_{hea}.

After the above steps, we can obtain the health probability of each individual. However, it is still difficult to decide which intervention action should be chosen. Therefore, we will introduce our HRLI model to solve this problem.

2.3 The Formulation of MDPs

In real life, both residences and individuals would take measures during an epidemic outbreak. Inspired by this, we model the residence and each individual as two agents and their decision making processes as two MDPs, in order to find an optimal intervention strategy.

2.3.1 MDP for Residence. We first define a *risk level* according to the number of infected people (N_{res}) in each residence. Table 1 shows the specific information of the *risk level*.

Table 1: The risk level of a residence.

Infected number	Risk level	Meaning
N_{res}=0	1	No-risk
0<$N_{res} \leq 10$	2	*Low-risk*
10<$N_{res} \leq 50$	3	*General-risk*
50<$N_{res} \leq 100$	4	*Medium-risk*
100<$N_{res} \leq 500$	5	*High-risk*
500<N_{res}	6	Serious-risk

In our model, each residence can use information of itself to adjust the risk level dynamically. Here, we show the specific definition of MDP for each agent (residence).

State S^{res} : The state of a residence is composed of 2 features, including $\frac{N_h^{inf}}{N_h}$, $\frac{\sum_{n=1}^{N_h^{hea}} \hat{p}_n}{N_h^{hea}}$, where N_h^{inf}, N_h and N_h^{hea} denote the number of infected (symptomatic and critical) people, the number of all people and the health people (except for the discovered patients) in the h-th residence, respectively.

Action A^{res}: The action includes three measures: raising the risk level, reducing the risk level and keeping the risk level unchanged.

Reward R^{res}: The reward can be expressed as as follows

$$R^{res} = -(e^{\frac{I^{res}}{500}} + e^{\frac{Q^{res}}{10000}}) \tag{4}$$

where I^{res} denotes the additional number of infected people on each day, $Q^{res} = N^{res}_{hos} + 0.5N^{res}_{iso} + 0.3N^{res}_{qua} + 0.2N^{res}_{con}$ denotes the weighted sum of the number of people who are intervened on each day, and N^{res}_{hos}, N^{res}_{iso}, N^{res}_{qua} and N^{res}_{con} denote the additional number of hospitalized, isolated, quarantined and confined people, respectively.

2.3.2 MDP for Individual. The specific definition of MDP for each individual is showed as follows:

State S^{ind}: The state of an individual is composed of 11 features, including *the risk level of its residence, its health probability, intervention state, infection state, the total number of infected people, the total number of hospitalized people, the total number of isolated people, the total number of quarantined people, the total number of confined people, the total number of stranger contacts and the total number of acquaintance contacts.*

Action A^{ind}: For each individual, the action includes three probabilities: $< p_1, p_2, p_3 > (0 \leq p_1 \leq p_2 \leq p_3 \leq 1)$. They meet the following rules in the Table 2.

Table 2: The relationship between intervention actions and health probability.

The health probability	Intervention actions
$0 \leq \hat{p} \leq p_1$	*no intervene*
$p_1 \leq \hat{p} \leq p_2$	*confine*
$p_2 \leq \hat{p} \leq p_3$	*quarantine*
$p_3 \leq \hat{p} \leq 1$	*isolate*

Reward $R^{ind}, \hat{R}^{ind}$: The long-term reward R^{ind} of 60 days can be expressed as the negative of the evaluation metric $Score$. However, the epidemiological model is a sparse reward space and the short-term reward $\hat{R}^{ind}$ of each day is not defined. Therefore, an theoretical concept in communicable disease epidemiology, *effective reproduction ratio* R_0, can be used to define the reward value of each day. R_0 denotes the expected number of secondary cases of an infection. If R_0 is greater than one, a newly introduced infection may lead to a large epidemic in a completely susceptible population. If R_0 is less than one, the total size of a newly introduced outbreak will remain small [1]. By combining long-term reward R^{ind} with short-term reward $\hat{R}^{ind}$, the reward of the individual can be expressed as follows:

$$\begin{aligned} R^{ind} &= -Score \quad if \quad day = 60 \\ \hat{R}^{ind} &= -1 \quad if \quad day < 60 \; and \; R_0 > 1 \\ \hat{R}^{ind} &= 1 \quad if \quad day < 60 \; and \; R_0 \leq 1 \end{aligned} \tag{5}$$

2.4 The HRLI Model

Figure 1 shows the flow chart of of HRLI model. In this model, the residence is designed as the top-level structure to adjust the *risk level* dynamically. Each individual is designed as the bottom-level structure, in order to handle each individual precisely and take the most reasonable intervention action.

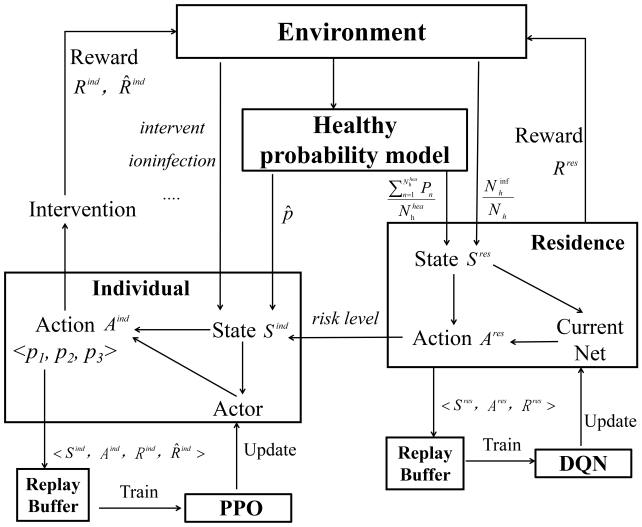

Figure 1: The flowchart of the HRLI model.

Specifically, the health probability model is first used to calculate the current health probability of the 10K-people. Then, the residences and individuals use DQN [2] and PPO [3] to learn a intervention strategy, respectively. We take 60 days as an episode and keep training the model until the reward converges.

3 CONCLUSIONS AND FUTURE WORK

Although the preliminary model has achieved some improvements, it still could not achieve the optimal strategy. In the future, we will continue to improve it from the following aspects:

- The risk level of a residence is the most important feature to coordinate the decision making processes between the two layers. Therefore, more precise definition and calculation of the risk level is essential.
- The definition of MDP needs to be adjusted to get more reasonable reward and state.
- More collaborative mechanisms can be added. For example, when taking measures for the overall risk level, the information of the neighboring residence can be considered.

REFERENCES

[1] K. Dietz. [n.d.]. *Reproduction Number*. American Cancer Society.
[2] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).
[3] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
[4] Unkown. [n.d.]. https://hzw77-demo.readthedocs.io/en/round2/simulator_modeling.html.
[5] Unkown. [n.d.]. https://prescriptive-analytics.github.io/.