Epidemic Control Based on Reinforcement Learning Approaches

Yuanshuang Jiang
UCAS
SIAT, CAS
JD.com
Shenzhen, China
ys.jiang@siat.ac.cn

Linfang Hou
JD.com
Beijing, China
houlinfang@jd.com

Yuxiang Liu
UCAS
SIAT, CAS
Shenzhen, China
yx.gao@siat.ac.cn

Zhuoye Ding
JD.com
Beijing, China
dingzhuoye@jd.com

Yong Zhang
SIAT, CAS
Shenzhen, China
zhangyong@siat.ac.cn

Shengzhong Feng
NSCCSZ
Shenzhen, China
fengsz@nsccsz.cn

ABSTRACT

In this work, we introduce an RL (Reinforcement Learning) algorithm to optimize the mobility intervention strategy in order to control the epidemic spreading. The performance of the proposed method is evaluated via interacting with the simulator provided by PAPW 2020 [1]. In the simulated environment, physical condition of each individual is dynamically classified into 5 cases. We developed different intervention strategies for different physical conditions. Our method help us win the first place in the competition.

CCS CONCEPTS

• **Applied computing** → **Health informatics**; **Multi-criterion optimization and decision-making**.

KEYWORDS

Reinforcement learning, Epidemic control

ACM Reference Format:
Yuanshuang Jiang, Linfang Hou, Yuxiang Liu, Zhuoye Ding, Yong Zhang, and Shengzhong Feng. 2020. Epidemic Control Based on Reinforcement Learning Approaches. In *PAPW '20: Workshop on Prescriptive Analytics for the Physical World, August 24, 2020, San Diego, California, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/xx.xxxx/xxxxxxx.xxxxxxx

1 INTRODUCTION

At present, COVID-19 is sweeping the world. The latest data from the World Health Organization shows that as of 10:00 on July 23 (Central European Time), the cumulative number of confirmed cases has reached about 15,012,731[6]. Under such a background, the Workshop of Prescriptive Analytics for the Physical World (PAPW 2020) organized a challenge of designing mobility intervention strategies for epidemics[5]. Participants are required to design effective mobility intervention strategies to contain epidemics. The performance of the strategy is evaluated by an epidemic simulator provided by the competition organizer. Specifically, the score of the applied intervention strategy is defined as follows.

$$\text{Score} = \exp\left\{\frac{I}{\theta_I}\right\} + \exp\left\{\frac{Q}{\theta_Q}\right\},$$

where I is the total number of infected people; Q is the weighted sum of times that each of the five intervention operations (i.e. *none*, *confine*, *quarantine*, *isolate* and *hospitalize*) is performed; θ_I and θ_Q are predefined scaling factors. Finally, the performance of a strategy is evaluated by the average score over multiple runnings with five distinct scenarios [5].

Reinforcement learning is essentially a sequential decision processes, and has gained great successes in many applications during the last decades [2, 4, 7]. It is natural that we design RL algorithms for this challenge, as the epidemic control strategy is actually a sequence of decisions. The epidemic situation and the interventions can be modeled as RL states and actions, respectively.

The main contribution of our method includes

- proposing a reinforcement learning approach to contain epidemics, and learning intervention strategies for different physical conditions.
- introducing an auxiliary-strategy-based learning method, where the strategy learned for those who got infected by strangers, helps deciding what intervention actions should be applied to all infected individuals.

2 METHOD

For an individual v, we use $p_{acq}(v)$ to denote the probability that v is infected via the contacts with the acquaintances, and use $p_{str}(v)$ to denote the probability that v is infected via the contacts with the strangers.

At first, we performed some basic rules. For recovered individuals and those infected ones who are in the hospital, none of the interventions will be performed. Similarly, for individuals who has a zero probability for both acquaintance infection and stranger infection, none of the interventions will be performed. The remaining individuals are further divided into 5 cases, and the corresponding intervention strategies are learned respectively. Furthermore, we

PAPW '20, August 24, 2020, San Diego, California, USA

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

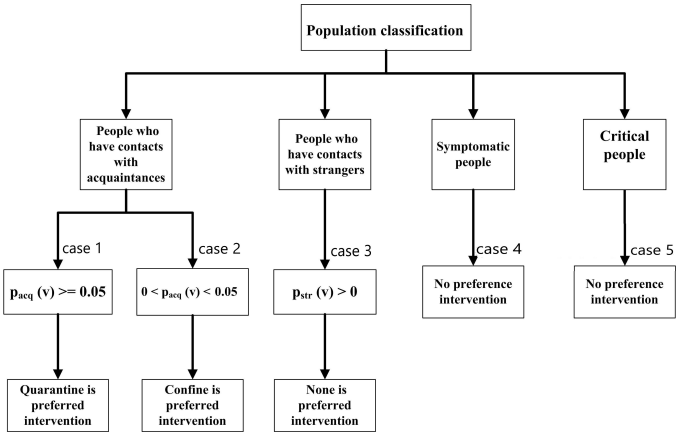

Figure 1: Population classification.

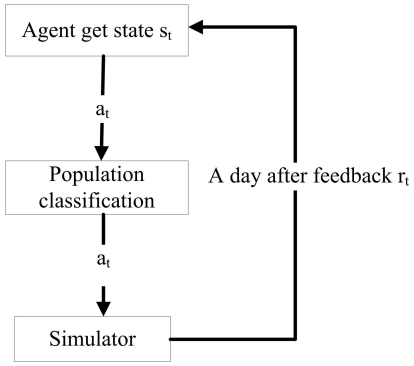

Figure 2: The basic flow of the method.

restrict preferred intervention choices according to the environmental simulator and effective experience in the real world , as shown in Figure 1. Detailed description are listed as follows.

- Case 1. For individuals with $p_{acq}(v) \geq 0.05$, we believe that they have high-infected-risks. Accordingly, *quarantine* is our preferred intervention in the learning process.
- Case 2. For individuals with $0 < p_{acq}(v) < 0.05$, *confine* is preferred.
- Case 3. For individuals with $p_{str}(v) > 0$, *none* is set as the preferred intervention in the learning process.
- Case 4. For individuals that have shown symptoms without severe illness, there is no preference among intervention choices.
- Case 5. For individuals that have shown symptoms with severe illness, we do not have preferred interventions.

Then, We conduct RL algorithm to learn intervention strategy for each case. Together, Figure 1 and Figure 2 demonstrate our basic ideas. As for our RL setting, we define *states*, *actions* and *rewards* as follows.

States. We have considered the joint-state of all the 10,000 individuals in the simulation. For each individual, the state consists of the following 8 dimensions.

(1) The individual's infection state, i.e. state 1/3/4/5 according to the simulator API [5].
(2) The interference state of this individual, i.e. state 0/1/2/3/4/5 according to the API.
(3) The aggregated infection probability within 5 days. According to the API, there are acquaintance infection probability and stranger infection probability. So we have 2 dimensions here, i.e. acquaintance infection state, stranger infection state.
   (a) Acquaintances infection state: the state is set to 2 if the aggregated probability is greater or equal to 0.05, and 1 if the aggregated probability is between 0 and 0.05. Otherwise, the state is 0.
   (b) Stranger infection state: the state is 1 if the aggregated probability is greater than 0; otherwise the state is 0.
(4) Time information.
(5) Day type: weekday or weekend.
(6) The number of contacts with acquaintances.
(7) The number of stranger contacts.

Action. The action space has 6 dimensions, The first 5 dimensions indicate the preference of each intervention choice, i.e., {*none*, *confine*, *quarantine*, *isolate*, *hospitalize*}, for different population cases after classification, i.e., {*case 1*, *case 2*, *case 3*, *case 4*, *case 5*}. The last dimension indicate the duration of performing an intervention. We demonstrate our action design in Figure 3.

With the multi-level interventions, the preference of an specified intervention is reflected by associating it with a wider value range. First, we reckon that different cases have different levels of priority. For example, when an individual belongs to case 1 and case 3 at the same time, priority should be given to case 1 because the population in case 1 is the high-risk group. Second, we reckon that different interventions have different levels of priority. For example, when an individual belongs to different cases but with the same priority level at the same time, such as case 2 and case 3, the intervention with higher priority level is performed. The priorities of the interventions is defined as *isolate* > *quarantine* > *confine* > *none*.

Values	0 - 5	6 - 11	12 - 17	18 - 23	24 - 29
Case 1	none	confine	quarantine	quarantine	isolate
Case 2	none	confine	confine	quarantine	isolate
Case 3	none	none	confine	quarantine	isolate
Case 4	none	confine	quarantine	isolate	hospitalize
Case 5	none	confine	quarantine	isolate	hospitalize

Figure 3: Values of the first 5 dimensions of RL action.

As for the last dimension, its value is ranged from 1 to 30. In particular, we found that the resulting strategy is slightly better if the action duration is always set to 1. However, this modification will cost a longer time period to learn an efficient strategy.

Reward. The reward of an action is defined as the incremental part of the cumulative scores between consecutive days.

To solve this RL problem, we use Proximal Policy Optimization (PPO) method [3], which execute multiple steps of mini-batch gradient decent updates by apply the experience in each iteration.

3 EXPERIMENTAL RESULTS

We have trained 3 models for 5 scenarios, especially for scenario 2, scenario 3, and scenario 4. In the evaluation phase, the method of scenario 2 was directly applied to scenario 1, 2, and 5. Table 1 show the best scores after training. We save the best training results every time for evaluation.

Table 1: The best scores during agent training

scenario 2	scenario 3	scenario 4
2.39	2.35	6.40

Through the analysis of the results of multiple scenarios, we found an interesting pattern: the learned strategies with high scores prefer to perform isolation and confine, while doing very little or even no quarantine intervention and hospitalise intervention. Therefore, we estimate the expectation of cumulative scores by conducting single intervention all the time, and find out that removing hospitalization choice resulting a better score. So we switch hospitalization to isolation for 30 days, because patients can recover themselves after 15-30 days, and the cost is lower than that of 15 days in the hospital. Experimental results show that our method works, and we should speculate that the removal of quarantine may also bring some improvement. Table 2 shows the final evaluation score.

Table 2: The final evaluation score

scenario 1	scenario 2	scenario 3	scenario 4	scenario 5
2.5	2.59	2.44	6.67	2.42

4 CONCLUSIONS

In this paper, we propose a hybrid method to learn epidemic control. We divide the population into 5 groups, and then learn effective strategies via RL algorithm for each case.

REFERENCES

[1] PAPW 2020. 2020. PAPW 2020 CFP. [n.d.]. ([n. d.]). https://prescriptive-analytics.github.io/challenge-cfp/index.html
[2] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. 2016. End-to-End Training of Deep Visuomotor Policies. *J. Mach. Learn. Res.* 17 (2016), 39:1–39:40.
[3] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *CoRR* abs/1707.06347 (2017). arXiv:1707.06347
[4] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of Go with deep neural networks and tree search. *Nat.* 529, 7587 (2016), 484–489.
[5] PAPW 2020. 2020. Simulator. [n.d.]. ([n. d.]). https://https://hzw77-demo.readthedocs.io/en/round2/
[6] WHO July 23 (Beijing time) COVID-19 epidemic information. 2020. (2020). https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200723-covid-19-sitrep-185.pdf?sfvrsn=9395b7bf_2
[7] Bin Wu. 2019. Hierarchical Macro Strategy Model for MOBA Game AI. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. 1206–1213.