KDD CUP 2019: HUMANITY RL TRACK SOLUTION BY ALPHA

{ SUIQIAN LUO } GUAZI.COM

瓜子 二手车直卖网 Guazi.com

INTRODUCTION

Malaria is thought to have had the greatest disease burden throughout human history, while it continues to pose a significant but disproportionate global health burden. With 50% of the world's population at risk of malaria infection. Sub Saharan Africa is most affected, with 90% of all cases.
Through this KDD Cup | Humanity RL track competition we are looking for participants to apply machine learning tools to determine novel solutions which could impact malaria policy in Sub Saharan Africa. Specifically, how should combinations of interventions which control the transmission, prevalence and health outcomes of malaria infection, be distributed in a simulated human population.
This task is about policy learning for malaria control in Sub Saharan Africa. To be more specific, we need to apply novel algorithms to solve a sequential decision making task.
The final score is the median of reward scores from 10 instantiations. Each instantiation consists of 21 episodes. From each episode, the algorithm will get a reward score. Scores in previous episodes can be used in later episodes in this instantiation. After running 21 episodes, the maximum score will be chosen as the score of this instantiation. Each episode includes 5 actions and can be denoted by 10 real numbers in total.

ENVIRONMENT ANALYSIS

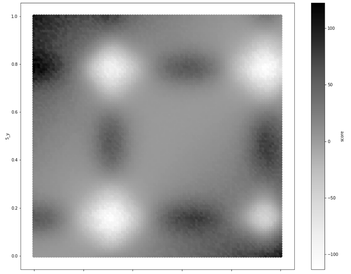

Figure 1: Fifth action

We investigate the environment directly and hope to find some meaningful patterns. We first consider a simple case, in which the first 4 actions are fixed at zero point.
In Figure 1, x and y axes represent the values of the fifth action and the color represents the corresponding reward score. The darker the color is, the higher the score is.
Next, we will move one step ahead and investigate the interaction between the last two actions.
In Figure 2, x and y axes represent the values of the fourth action, and each small grid is the visualization of the fifth action and score. First, with the increase of x in the fourth action, the pattern is scaled proportionally along the same direction, and so as y. Second, the background pattern looks like the basic pattern, which is also shown in the bottom-left.
Based on the two findings, we hypothesize that the reward score can be decomposed into five parts. Each part of score is only affected by the previous and current actions. We conduct extensive random trials and all results confirm our assumption.
For the second environment, we use the same assumption in previous phase. For the first action, we find that there are distinct boundaries forming 8 by 8 grids. For the interaction between two adjacent actions, the patterns are hard to summarize.

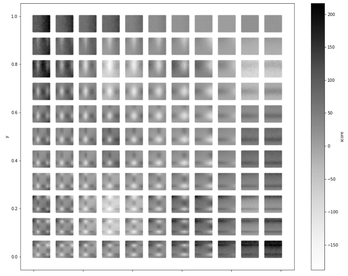

Figure 2: Fourth and fifth action

MODEL

Based on our findings in the first two phases, we summarize some points below.
We think that the gradient-based methods may be not effective. We can not determine whether the value space is composed by grids or not.
Because we only have 21 chances to optimize the reward score, we do not have enough chances to detect and use the patterns if any.
Therefore, we come up with our final solution.
We use the random search, which is a very effective gradient-free black-box method, in a two-stage form.
We separate 21 episodes into two groups.
In the first group, we use 6 episodes to randomly choose actions. Then we keep the best action. In the second group, we use 15 episodes for random adjustments. In each 5 episodes, we adjust 5 actions in order. And we do this for 3 rounds. In each adjustment, if we get a higher score, we will accept the change.
Here are some tricks about our solution. Because of the limitation on number of trials, we prefer to explore larger areas of searching space. To achieve this, we restrict the number choices in these five numbers. Besides, we avoid to choose the same action.
The code is public:
https://github.com/luosuiqian/submission