Can Reinforcement Learning Lead to Healthy Life?: Simulation Study Based on User Activity Logs

Masami Takahashi, Masahiro Kohjima, Takeshi Kurashima, Hiroyuki Toda (NTT)

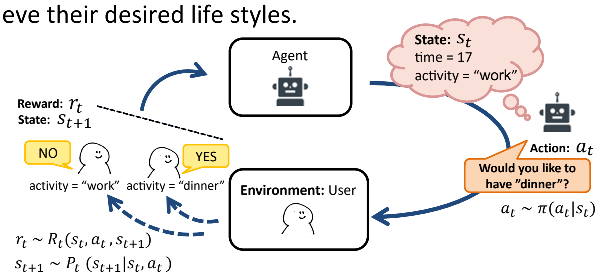
Summary

We propose an automatic intervention method based on RL to help users achieve their health goals (e.g., sleep at 10:00 p.m. to get enough sleep). Our method estimates a user model (transition probability) and then computes the optimal intervention strategies given the user model and goals. We construct the user model based on real activity data and confirm the effectiveness of the proposed RL-based interventions.

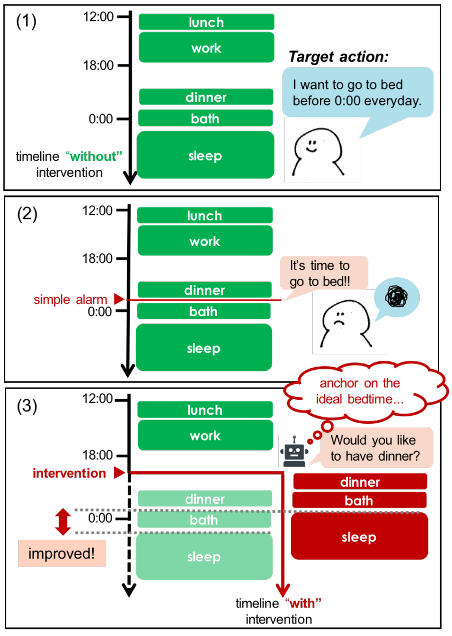
Introduction

Motivation: A challenging part of realizing the application that leads to a healthier life is the need for planning, i.e., considering the user's health goal, providing intervention at the appropriate timing to help the user achieve the goal. The reinforcement learning (RL) approach is well suited to this type of problem since RL makes decisions based on planning that consider the effect of a current decision on the future.

We propose an automatic intervention method based on RL and investigate the effects of RL-based intervention to help users achieve their desired life styles.

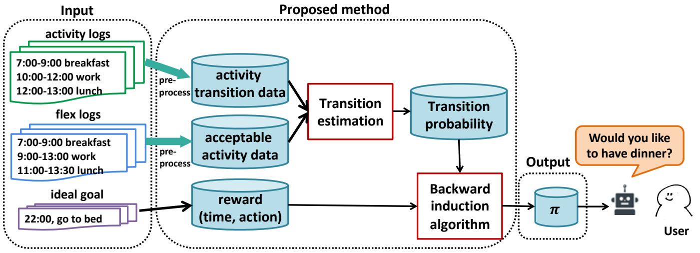


State, action and reward in this RL problem correspond to the user's activity, app's intervention, and user's goal.



Example: a user who wants to go to bed earlier than usual.
(1) The user specifies the goal.
(2) It is difficult to respond to a simple intervention if the activities that must be performed prior to sleeping have not been completed.
(3) Our app encourages the user to complete prior dependent actions with the goal of sleep.

Proposed Method



Transition Probability Estimation: We denote the model of the transition probability as $P_t^\theta(s_{t+1} = j | s_t = i, a = k)$ to P_{tijk}^θ.

Using the model gives the likelihood function of the data as

$$P(\mathcal{D}^{tr}|\theta) = \prod_{t=1}^{T} \prod_{i,j \in \mathcal{S}} (p_{tij|\mathcal{A}|}^\theta)^{N_{tij}}$$

The likelihood function of the acceptable activity data can be written as

$$P(\mathcal{D}^{apt}|\theta) = \prod_{t=1}^{T} \prod_{i,j \in \mathcal{S}} (p_{tijj}^\theta)^{\beta M_{tj}} (1 - p_{tijj}^\theta)^{\{(1-\beta) M_{tj} + (L - M_{tj})\}}$$

Taking the negative logarithm of the above two likelihood functions yields the following objective function:

$$\mathcal{L}(\theta) = -\log P(\mathcal{D}^{tr}|\theta) - \gamma \log P(\mathcal{D}^{apt}|\theta) + \Omega(\theta),$$

θ is estimated by optimizing this objective $\hat{\theta} = \arg min_\theta \mathcal{L}(\theta)$.

Backward Induction Algorithm: Given the estimated transition probability and reward function, our system outputs the optimal policy by value iteration.

Algorithm 1 Backward Induction Algorithm for Finite-Horizon Entropy-regularized RL

Input: $\mathcal{P}$: transition probability, $\mathcal{R}$: reward function, α: hyperparameter

Output: $\{Q_t^*\}_t, \{V_t^*\}_t$: value function, $\{\pi_t^*\}_t$: policy

1: Set $t \leftarrow T$ and $V_T(s) = 0$ for all $s \in \mathcal{S}$.

2: Set $t \leftarrow t - 1$

3: Compute $Q_t(s, a)$ following

$$Q_t^{\pi^*}(s, a) = \mathbb{E}_{s' \sim \mathcal{P}_t(s'|s,a)}[\mathcal{R}_t(s, a, s') + V_{t+1}^{\pi^*}(s')]$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

4: Compute $V_t(s)$ for all $s \in \mathcal{S}$ following

$$V_t^{\pi^*}(s) = \alpha \log \sum_{a'} \exp\left(\alpha^{-1} Q_t^{\pi^*}(s, a')\right).$$

5: Compute $\pi_t(a|s)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$ following

$$\pi_t^*(a|s) = \exp\left(\alpha^{-1}\{Q_t^{\pi^*}(s, a) - V_t^{\pi^*}(s)\}\right).$$

6: If $t = 0$, stop. Otherwise, return to step 2.

Experiment

In the simulations, we calculated the mode start time of bedtime from the collected activity data, and set the goal time for each participant: (a) mode time, (b) mode time -1 hours, and (c) mode time -2 hours. We compared the average return by the proposed method with the baselines (random intervention and alarm settings). The results show our method attained the highest rewards.

TABLE II: Average reward values of all participants (proposed method, random, one time). Larger is better.

	proposed	random	one time
mode	59.23(±19.38)	-22.17(±23.51)	55.70(±20.02)
mode-1	23.07(±23.54)	-53.47(±19.62)	20.60(±22.23)
mode-2	12.77(±20.14)	-58.88(±15.32)	11.65(±19.30)