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.

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.

Proposed Method
Input
activity logs
outputs
Transition probability
Backward Induction Algorithm
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: \( P \): transition probability, \( R \): reward function, \( \alpha \): hyperparameter
Output: \( \{ Q^*_t \}, \{ \pi^*_t \} \): value function, \( \{ \pi^*_t \} \): policy
1: Set \( t \leftarrow T \) and \( V_T(s) = 0 \) for all \( s \in S \).
2: Set \( t \leftarrow t - 1 \).
3: Compute \( Q_t(s, a) \) following
\[
Q_t(s, a) = \mathbb{E}_{s' \sim P_t(s,a)}[R_t(s, a, s') + V_{t+1}(s')] \]
for all \( s \in S \) and \( a \in A \).
4: Compute \( V_t(s) \) for all \( s \in S \) following
\[
V_t(s) = \alpha \log \sum_{a'} \exp(\alpha^{-1}Q^*_t(s, a')) \]
5: Compute \( \pi_t(a|s) \) for all \( s \in S \) and \( a \in A \) following
\[
\pi_t(a|s) = \exp\left(\alpha^{-1}Q^*_t(s, a) - V^*_t(s)\right) \]
6: If \( t = 0 \), stop. Otherwise, return to step 2.

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

<table>
<thead>
<tr>
<th></th>
<th>proposed</th>
<th>random</th>
<th>one time</th>
</tr>
</thead>
<tbody>
<tr>
<td>mode</td>
<td>59.23±19.38</td>
<td>-22.17±23.51</td>
<td>55.70±20.02</td>
</tr>
<tr>
<td>mode-1</td>
<td>23.07±23.54</td>
<td>-53.47±19.62</td>
<td>20.60±22.23</td>
</tr>
<tr>
<td>mode-2</td>
<td>12.77±20.14</td>
<td>-58.88±15.32</td>
<td>11.65±19.30</td>
</tr>
</tbody>
</table>

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.