An Application of Inverse Reinforcement Learning to Medical Records of Diabetes Treatment

H. Asoh, M. Shiro, S. Akaho, T. Kamishima
(National Institute of Advanced Industrial Science and Technology)
K. Hasida (The University of Tokyo)
E. Aramaki (Kyoto University)
T. Kohro (The University of Tokyo Hospital)

September 23, 2013
ECMLPKDD2013 Workshop on Reinforcement Learning with Generalized Feedback: Beyond Numeric Rewards
Introduction

- Long-term process of medical treatments for chronic diseases can be considered as interactions between patients and doctors.
- We are exploiting a MDP to model the long term interaction processes of disease treatment.

- State: patient's Status (result of examinations)
- Action: medical treatments

<table>
<thead>
<tr>
<th>State</th>
<th>Action</th>
<th>Reward</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1 = \text{normal}$</td>
<td>$A_1 = \text{RapidIns}$</td>
<td>1</td>
</tr>
<tr>
<td>$S_1 = \text{medium}$</td>
<td>$A_2 = \text{TDZ + SU}$</td>
<td>0</td>
</tr>
</tbody>
</table>

doctor

patient

$A_1 = \text{RapidIns}$

$A_2 = \text{TDZ + SU}$

reward = 0

reward = 1

reward = 0
Introduction

- Using the estimated MDP, we can
  - Predict progression of treatments
  - Evaluate value of patient's states
  - Evaluate value of doctor's actions

- Related Work
  - Optimal timing of living-donor liver transplantation [Alagoz+ 2004]
  - Optimal time to initiate HIV therapy [Shechter+ 2008]
  - Modeling treatment process of ischemic heart disease [Haskrecht+ 2000]
Introduction

- We focus on the process of controlling blood glucose level for type 2 Diabetes patients
  - Large social impact
    - 8.3% of the U.S. population (2011)
    - 11.6% of the total health care expenditure in the world for 2030
  - Lead to very serious complications including heart diseases
Data

- Records of patients cared at the University of Tokyo Hospital for their heart diseases (around 3,000 patients)
- We extracted patients with periodical visits
  - Interval between visits was more than 15 days and less than 75 days (around 1 month)
  - Longer than 24 visits
- 801 patients were extracted
  - Minimum length: 25 visits (around 2 years)
  - Maximum length: 124 visits (over 10 years)
Data

- **State**: value of Hemoglobin-A1c (HbA1c)

<table>
<thead>
<tr>
<th>Level</th>
<th>Normal</th>
<th>Medium</th>
<th>Severe</th>
</tr>
</thead>
<tbody>
<tr>
<td>HbA1c</td>
<td>&lt; 6.0</td>
<td>6.0 - 8.0</td>
<td>&gt; 8.0</td>
</tr>
</tbody>
</table>

- **Action**: pharmaceutical treatments

  - Alpha-Glucosidase Inhibitor (αGI)
  - Biganaides (BG)
  - DPP4 Inhibitor (DPP4)
  - Insulin (Ins)
  - Rapid-Acting Insulin Secretagogue (RapidIns)
  - Sulfonyurea (SU)
  - Thiazolidinedion (TDZ)

  7 types of drug
  38 combination patterns
e.g. αGI+DPP4+SU
Data

- Reward: No reward value in the data
- We assumed a simple reward: e.g.
  - if state == "normal" reward = 1
  - else reward = 0

Example of an episode

<table>
<thead>
<tr>
<th>time</th>
<th>state1</th>
<th>action</th>
<th>state2</th>
<th>reward</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000/1/1</td>
<td>medium</td>
<td>TDZ</td>
<td>medium</td>
<td>0</td>
</tr>
<tr>
<td>2000/2/3</td>
<td>medium</td>
<td>αGI+DPP4</td>
<td>normal</td>
<td>1</td>
</tr>
<tr>
<td>2000/3/1</td>
<td>normal</td>
<td>no action</td>
<td>normal</td>
<td>1</td>
</tr>
<tr>
<td>2000/4/5</td>
<td>normal</td>
<td>no action</td>
<td>medium</td>
<td>0</td>
</tr>
</tbody>
</table>
State Values and Action Values

- Estimate MDP parameters using Data
- Evaluate state/action values [Asoh+ 2013]
Issue

- Reward values are not in the data
- We assumed simple reward function based on the purpose of the analysis

Question: What kind of reward the doctors have in their mind?

- Applying IRL to the medical records
Algorithms of IRL

- Linear programing [Ng+ 2000]
- Quadratic programing [Abbeel+ 2004]
- Bayesian IRL [Ramachandran+ 2007]
- Extension of the Bayesian IRL [Rothkopf+ 2011]
Bayesian IRL [Ramachandran+ 2007]

- Known MDP environment
- Finite discrete state space
- Reward depends only on state
  - Reward function $R$ is represented as a vector
- Probabilistic generative model of experts' behavior (state-action pairs)

$$Pr_{\chi}((s, a) | R) = \frac{1}{Z} e^{\alpha \chi Q^*(s, a; R)}$$
Bayesian IRL

- A sequential observation of experts' behaviours
  \[ O_\chi = \{(s_1, a_1), (s_2, a_2), \ldots, (s_k, a_k)\} \]
- Posterior probability of reward vector
  \[ Pr_\chi(R|O_\chi) = \frac{1}{Z'} e^{\alpha E(O_\chi;R)} P_R(R) \]
  \[ E(O_\chi;R) = \sum_{i=1}^{k} Q^*(s_i, a_i; R) \]
Bayesian IRL

- MCMC Sampling from the posterior distribution or reward vector
- Policy Walk algorithm
  ✓ Combining Policy Iteration for MDP and Metropolis-Hastings Algorithm

\[ R = (R_{normal}, R_{medium}, R_{severe}) \]
\[ R_{normal} + R_{medium} + R_{severe} = 1 \]
Result

$R_{medium}$
Discussion

Converged to $R^*= (0.01, 0.98, 0.01)$

<table>
<thead>
<tr>
<th>$R$</th>
<th>$(0.98, 0.01, 0.01)$</th>
<th>$(0.01, 0.98, 0.01)$</th>
<th>$(0.01, 0.01, 0.98)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log-likelihood</td>
<td>-159878</td>
<td>-143568</td>
<td>-162928</td>
</tr>
</tbody>
</table>

Possible causes of the counter-intuitive result

- Many of the patients were already in the “medium” state when they came to the hospital

<table>
<thead>
<tr>
<th>State</th>
<th>normal</th>
<th>medium</th>
<th>severe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rel. Frequency</td>
<td>0.178</td>
<td>0.65</td>
<td>0.172</td>
</tr>
</tbody>
</table>

- keeping the patients’ state at “medium” may be the best-effort target of doctors.
Discussion

- Other possible causes of the counter-intuitive result
  - the MDP model is too simple to model the decision-making process of doctors
  - assuming that the reward value depending only on the current state is too simple
  - heterogeneity of doctors and patients is not properly considered
Discussion

Comparison between

✓ the doctors' policy and
✓ the optimal policy under the estimated reward value $R^*$

<table>
<thead>
<tr>
<th>State</th>
<th>normal</th>
<th>medium</th>
<th>severe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal policy under $R^*$</td>
<td>BG+SU</td>
<td>$\alpha$GI+BG+SU+TDZ</td>
<td>DPP4</td>
</tr>
<tr>
<td>Doctors' policy</td>
<td>SU</td>
<td>SU</td>
<td>SU</td>
</tr>
<tr>
<td></td>
<td>$\alpha$GI</td>
<td>BG+SU</td>
<td>BG+SU</td>
</tr>
<tr>
<td></td>
<td>TDZ</td>
<td>$\alpha$GI</td>
<td>BG</td>
</tr>
</tbody>
</table>
Summary

- The process of medical treatment for diabetes was modeled with a MDP

- A Bayesian IRL algorithm was applied to the MDP environment

- The result was counter-intuitive
  - Reward for “medium” state of patient is high
Future Study

- Detailed validation of the result
  - Using different algorithms
  - Using different state representations

- More complex decision-making model may be necessary
  - Introducing medical knowledge regarding pharmaceutical treatments
  - Consulting guidelines for treatment
  - Detailed modeling of physicians’ therapeutic decisions [Toussi+ 2009]
Thank you, and we would like to learn more from your “non numeric” feedbacks!