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

Hideki Asoh[1], Masanori Shiro[1] Shotaro Akaho[1], Toshihiro Kamishima[1], Koiti Hasida[1], Eiji Aramaki[2], and Takahide Kohro[3]

[1] National Institute of Advanced Industrial Science and Technology, AIST Tsukuba Central 2, Tsukuba, Ibaraki 305-8568, Japan,
h.asoh@aist.go.jp, shiro@ni.aist.go.jp, s.akaho@aist.go.jp, mail@kamishima.net, hasida.k@aist.go.jp

[2] Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
eiji.aramaki@gmail.com

[3] The University of Tokyo Hospital, Hongo, Bunkyo-ku, Tokyo 113-8655, Japan
kohro-tk@rc5.so-net.ne.jp

Abstract. It is an important issue to utilize large amount of medical records which are being accumulated on medical information systems to improve the quality of medical treatment. The process of medical treatment can be considered as a sequential interaction process between doctors and patients. From this viewpoint, we have been modeling medical records using Markov decision processes (MDPs). Using our model, we can simulate the future of each patient and evaluate each treatment. In order to do so, the reward function of the MDP should be specified. However, there is no explicit information regarding the reward value in medical records. In this study we report our results of applying an inverse reinforcement learning (IRL) algorithm to medical records of diabetes treatment to explore the reward function that doctors have in mind during their treatments.

Keywords: inverse reinforcement learning, medical records, diabetes

1 Introduction

The process of medical treatment can be considered as interaction between doctors and patients. Doctors observe the state of patients through various types of examination and accordingly choose the proper treatment, which causes a change of state in the patient. In particular, lifestyle-related chronic diseases such as diabetes mellitus, hypertension, hyperlipidemia, and vascular disorders slowly progress and the interaction can last for several years.

Conventional statistical analysis of medical data mainly evaluates the impact of a single treatment or a single factor on the target events. Analyzing records of such long-term treatment to evaluate the costs and the benefits on the quality of life of the patient remains under investigation.

Aiming to address this issue, we exploited Markov decision processes (MDPs) to model the long-term process of disease treatment [1]. The parameters of the MDP were estimated by using patients' medical records, and We estimated the progression of the patients' state and evaluated the value of patients' states and doctors' treatments (actions) using the estimated mode. We also applied a type of reinforcement learning (RL) to get the optimal action selection rule (policy) depending on the state of the patient.

To evaluate the state and action values, and to conduct RL, setting an appropriate reward function according to the patient's state and the doctor's action is necessary. In our previous study [1], because the reward value is not included in the medical records, we assumed a simple reward function based on a doctor's opinion. However, the appropriateness of the reward function could not be validated.

In the recent studies on RL, inverse reinforcement learning (IRL) has been proposed to estimate the reward function of experts from their behavior data [2–6]. In this study, we apply a Bayesian IRL algorithm to the medical records and explore the reward function that are used by doctors.

The rest of this paper is organized as follows. In Section 2, model-based RL based on the MDP and IRL on the MDP are briefly introduced. In Section 3, the data used in this study is described. In Section 4, results of IRL are demonstrated and discussed, and Section 5 is for conclusions and remaining issues.

2 MDP, RL, and IRL

MDP is a common probabilistic generative model of time series data with sequential decisions and rewards [7–9]. A stationary MDP is composed of four elements: S, A, T, and R. S is a set of states, A is a set of actions, $T : S \times A \times S \to [0, 1]$ is a set of state transition probabilities, and $R : S \times A \times S \times \mathcal{R} \to [0, 1]$ is a set of probabilities of taking an immediate reward value for a state transition.

$$T(s, a, s') = Pr\{s_{t+1} = s' | s_t = s, a_t = a\}$$
$$R(s, a, s', r) = Pr\{r_{t+1} = r | s_t = s, a_t = a, s_{t+1} = s'\}$$

We also use the expected value of an immediate reward defined as follows:

$$\begin{aligned} R(s, a, s') &= E[r_{t+1} | s_t = s, a_t = a, s_{t+1} = s'] \\ &= \sum_r R(s, a, s', r)\, r \end{aligned}$$

For an MDP, we can define a policy $\pi : S \times A \to [0, 1]$, as a set of probabilities of choosing action $a \in A$ at state $s \in S$. Under the policy π, an episode (path) $s_0, a_1, s_1, r_1, a_2, s_2, r_2, ..., a_T, s_T, r_T$ can be sampled and we can compute the value of the cumulative reward $v = \sum_{t=1}^{T} \gamma^t r_t$ along an episode. Here $\gamma \in [0, 1]$ is a discount factor for future rewards.

For each state $s \in S$, the expectation of the cumulative reward under the policy π can be computed as follows:

$$V^{\pi}(s) = E\left[\sum_{t=1}^{\infty} \gamma^t r_t \Big| s_0 = s, \pi\right].$$

Here $E[\ ]$ denotes taking expectation over the probability distribution of episodes. $V^{\pi}(s)$ is called the value of the state s under the policy π.

Similarly, the value of an action a at state s under the policy π can be defined as follows:

$$Q^{\pi}(s,a) = E\left[\sum_{t=1}^{\infty} \gamma^t r_t \Big| s_0 = s, a_1 = a, \pi\right].$$

The optimal policy for an MDP is a policy π^* that satisfies $V^{\pi^*}(s) \geq V^{\pi}(s)$ for all state $s \in S$ and for all policy π. The state values for the optimal policy satisfy the following Bellman optimality equation:

$$V^*(s) = \max_a \sum_{s'} T(s,a,s')\{R(s,a,s') + \gamma V^*(s')\}.$$

For a given MDP and policy, we can evaluate state values $V^{\pi}(s)$ and action values $Q^{\pi}(s,a)$. We also can solve the MDP to obtain the optimal policy π^* and evaluate the state and action values under the optimal policy by various types of RL algorithm.

The problem of IRL can be formulated as the problem of estimating the reward function of an expert agent who behaves optimally under an environment and the reward function [2]. Several IRL algorithms have been proposed so far. Ng and Russel proposed an algorithm using linear programing [2]. Abbeel and Ng used quadratic programing. They used the estimated reward function to mimic the behavior of the expert agent, and call the learning process "apprenticeship learning" [3]. Ramachandran and Emir formulated the IRL problem in Bayesian framework, and proposed an algorithm using Markov chain Monte Carlo sampling [4]. Rothkopf and Dimitrakakis formally reformulated the Bayesian IRL within the framework of preference elicitation [5]. They also extend their method in the context of multitask learning [6].

Among them, we exploited the Bayesian IRL algorithm proposed in [4] because of the simplicity of the method. In their formulation, based on the Markov assumption, the probability of generating a sequence of observations and actions $O = \{(s_1,a_1),(s_2,a_2),...,(s_k,a_k)\}$ under the reward function R is decomposed as

$$Pr(O|R) = \prod_{i=1}^{k} Pr((s_i,a_i)|R).$$

It is also assumed that the experts decide their action based on the following Boltzmann distribution, derived from the action value function $Q^*(s,a,R)$ under the optimal policy for R.

$$Pr((s_i,a_i)|R) = \frac{1}{C_i}\exp\{\alpha Q^*(s_i,a_i,R)\}.$$

Here α is a parameter that controls the variance of the actions of experts. When α is large, the variance becomes small. C_i is the normalization constant and

$$C_i = \sum_{s \in S, a \in A} \exp\{\alpha Q^*(s, a, R)\}.$$

Because this value does not depend on i, we will write C.

Hence, the probability of generating the observation and action sequence O can be written as follows:

$$Pr(O|R) = \frac{1}{Z} \exp\{\alpha \sum_i Q^*(s_i, a_i, R)\}.$$

Here $Z = C^k$ is the normalization constant. Combined with the prior distribution of the reward function R, the posterior probability of R under the observation and action sequence O can be computed using Bayes's theorem as

$$Pr(R|O) = \frac{1}{Z'} \exp\{\alpha \sum_i Q^*(s_i, a_i, R)\} Pr(R).$$

Here, Z' is another normalization constant.

Let S be the set of state values. In the following we assume that S is a finite set. We also assume that the value of the reward function R depends only on the state $s \in S$. Hence, R can be described by a $|S|$ dimensional reward vector $\boldsymbol{r}$.

Several candidates of the prior distribution of $\boldsymbol{r}$ have been proposed. The simplest one is the uniform distribution over $[R_{min}, R_{max}]^{|S|}$. Gaussian distribution, Laplace distribution, and beta distribution can also be used.

To compute the posterior distribution of $\boldsymbol{r}$, the Markov chain Monte Carlo sampling is exploited. The "Policy Walk" algorithm based on the Metropolis-Hastings algorithm will be shown below.

Because the optimal policy is invariant under the scaling of the reward vector $\boldsymbol{r}$, we assume that all elements r_i of $\boldsymbol{r}$ are in the interval $[0, 1]$ and $\sum_i r_i = 1$. This means that $\boldsymbol{r}$ is in a simplex in the $|S|$ dimensional space.

We quantize the range of $\boldsymbol{r}$ by a grid. Let the distance between neighboring nodes be δ and let $U = R^{|S|}/\delta$ be a set of nodes. Then, for a known MDP M, the Policy Walk algorithm can be written as follows:

1. Initialize the set of samples of $\boldsymbol{r}$ as $G := \phi$.
2. Randomly choose an initial reward vector $\boldsymbol{r}$ from the set U.
3. Using policy-iteration, obtain the optimal policy under the $\boldsymbol{r}$ as $\pi := \text{PolicyIteration}(M, \boldsymbol{r})$.
4. Repeat the following steps:
   (a) Randomly choose a $\boldsymbol{r}'$ from the neighbors of $\boldsymbol{r}$.
   (b) Evaluate action values $Q^\pi(s, a, \boldsymbol{r}')$ under the current policy π for all state-action pair (s, a).
   (c) When π is not optimal under the $\boldsymbol{r}'$

i. Update the policy using policy-iteration under $\boldsymbol{r}'$ by using the current policy π as the initial policy for the policy-iteration: $\pi' := \text{PolicyIteration}(M, \boldsymbol{r}'; \pi)$.
ii. Evaluate action values under π'.
iii. Compute $p = \frac{Pr(O|\boldsymbol{r}')}{Pr(O|\boldsymbol{r})}$, and update the reward vector and policy in the probability $\min(1, p)$ as $\boldsymbol{r} := \boldsymbol{r}'$ and $\pi := \pi'$.

(d) When π is optimal under the $\boldsymbol{r}'$
i. Compute $p = \frac{Pr(O|\boldsymbol{r}')}{Pr(O|\boldsymbol{r})}$, and update the reward vector and policy in the probability $\min(1, p)$ as $\boldsymbol{r} := \boldsymbol{r}'$ and $\pi := \pi'$.

(e) Add $\boldsymbol{r}$ to the set of samples G.

5. Output G.

3 Data: Medical Records of Diabetes Treatment

The data used in this study were extracted from the medical records of heart disease patients cared for by the University of Tokyo Hospital. After obtaining approval for our research plan from the Institutional Review Board at the University of Tokyo and the National Institute of Advanced Industrial Science and Technology, we prepared the data for this study. First, we anonymized the data, and then extracted the records to be analyzed. We focused on patients who periodically attended the hospital and underwent examinations and pharmaceutical treatments. The data after January 1, 2000 were used.

In this study we exclusively focus on the data related to diabetes. This is because diabetes is a typical chronic disease that frequently generates long-term treatment records. In addition, the disease is a very important factor in complications leading to heart diseases, and thus it is worth analyzing.

Among the examinations in the data, we focused the value of hemoglobin-A1c (HbA1c) for simplifying the analysis. The value of HbA1c was discretized into three levels (normal, medium, and severe) according to two threshold values (6.0 and 8.0) suggested by a domain expert.

For pharmaceutical treatments, we grouped the drugs according to their functions and identified patterns of combinations of drug groups prescribed simultaneously. The number of identified combination patterns that appeared in the data was 38. Name of the drug groups and their abbreviations are shown in Table 1.

An event in a sequence is a visit of a patient to the hospital, i.e., a combination of an examination and a pharmaceutical treatment. Sometimes, no drug is prescribed. The intervals between successive events are almost stable (about 30 days), but sometimes vary. For the collection of periodic visit sequences, when the interval is longer than 75 days, we cut the sequence at that point. Thereafter we collected sequences that were longer than 24 visits (about 2 years). The total number of extracted sequences was 801. The characteristics of the data are summarized in Table 2.

Drug Group	Abbreviation
Alpha-Glucosidase Inhibitor	αGI
Biguanaides	BG
DPP4 Inhibitor	DPP4
Insulin	Ins
Rapid-Acting Insulin	RapidIns
Sulfonylurea	SU
Thiazolidinedion	TDZ

Table 1. Drug groups for diabetes treatment and their abbreviations

Property	Value
Number of episodes	801
Total number of events	32,198
Maximum length of an episode	124
Minimum length of an episode	25

Table 2. Characteristics of the analyzed data

4 Experiment

4.1 Experimental conditions

Using the extracted episodes from the records of diabetes treatments, we first estimated the state transition probabilities T of the MDP and the policy π of doctors. The discretized HbA1c values consists of a set of state S. The combinations of medicines correspond to action set A. In the probability estimation, we exploited Laplace smoothing to avoid problems caused by the sparseness of data. We set the smoothing parameter equals to 1 as default.

We then applied the IRL algorithm introduced in Section 2. Because we discretized the observation (state) value into three levels, the reward function R can be described by a three-dimensional vector $\boldsymbol{r} = (R_{normal}, R_{medium}, R_{severe})$ We normalized the vector as $R_{normal} + R_{medium} + R_{severe} = 1$. The value of δ and α was set as $\delta = 0.01$ and $\alpha = 1$ respectively. As the prior distribution for the $\boldsymbol{r}$ we used the uniform distribution over the simplex.

We implemented the whole program using the statistical computing language R. For the policy-iteration in the Policy Walk algorithm, we used a package of R for RL which we are currently developing.

4.2 Result and Discussion

As the result of MCMC sampling, we got $\boldsymbol{r} = (0.01, 0.98, 0.01)$ with probability nearly equal to 1. Figure 1 shows a typical sampling sequence of R_{medium}. This means that the reward value for the "medium" state is the highest.

This result seems counter-intuitive because the aim of treatment is considered to move the patient into "normal" state. To confirm the result, we compared the log likelihood values of the observations under the reward vector (0.98 0.1, 0.1),

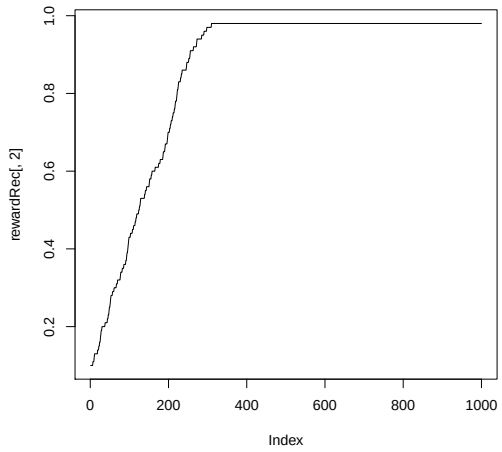

Fig. 1. Typical sample sequence of the reward value of "medium" states

(0.1, 0.98, 0.1), and (0.1, 0.1, 0.98). They are -159878, -143568, and -162928 respectively. Hence, the obtained reward vector has the maximum value.

An interpretation of the result is that our data is from patients and many of them were already in the "medium" state when they came to the hospital. The relative number of occurrences of the three states among the data is (0.178, 0.65, 0.172). Hence, keeping the patients' state at "medium" may be the best-effort target of doctors. Other possible reasons for the unexpected result are as follows:

- the MDP model is too simple to model the decision-making process of doctors,
- heterogeneity of doctors and patients is not properly considered.

We computed the optimal policy under the reward vector $\boldsymbol{r} = (0.01, 0.98, 0.01)$. Optimal actions for three states are "BG+SU", "aGI+BG+SU+TDZ", and "DPP4", respectively. We also evaluated the similarity between the doctors' policy and the optimal policy under the estimated reward function. The similarity is unfortunately not so high. This suggests that the assumption of the reward function depending purely on the patient's current state may be too simple.

5 Conclusions and Future Work

In this paper we reported our results of applying the Bayesian IRL to longitudinal medical records of diabetes patients. The results we have obtained so far are rather preliminary and many issues remain to be addressed.

The most essential problem is that the MDP that we exploited is very simple. The applicability of the Markov assumption should be confirmed first. In addition, we are considering some extensions of the MDP. The first one is introducing

the heterogeneity of doctors and patients by hierarchical modeling. Application of the multitask IRL algorithm by Dimitrakakis and Rothkopf [6] seems to be interesting. The second one is to extend the MDP to PODMP by introducing hidden state variables. The third one is introducing complex decision-making processes of doctors by considering longer histories of examinations and treatments. When we make the MDP more complex, the number of state values and actions will increase. Hence properly designing the state and the action spaces is also an important issue.

Analyzing long-term medical records of patients suffering from chronic diseases is beginning to be recognized as an important research issue, but it has not been investigated extensively. The approach shown in this study appears promising, and we would like to continue the investigation along these lines.

Acknowledgement: We thank anonymous reviewers for their constructive comments. This work was supported by JSPS KAKENHI Grant Number 23240043.

References

1. Asoh, H., Shiro, M., Akaho, S., Kamishima, T., Hasida, K., Aramaki, E., T.Kohro: Modeling medical records of diabetes using Markov decision processes. In: Proceedings of the ICML2013 Workshop on Roll of Machine Learning for Transforming Healthcare. (2013)
2. Ng, A.Y., Russell, S.: Algorithms for inverse reinforcement learning. In: Proceedings of the 17th International Conference on Machine Learning (ICML 2000). (2000) 663–670
3. Abbeel, P., Ng, A.Y.: Apprenticeship learning via inverse reinforcement learning. In: Proceedings of ICML 2004. (2004)
4. Ramachandran, D., Amir, E.: Bayesian inverse reinforcement learning. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence. (2007)
5. Rothkopf, C.A., Dimitrakakis, C.: Preference elicitation and inverse reinforcement learning. In: Proceedings of ECML-PKDD 2011. (2011)
6. Dimitrakakis, C., Rothkopf, C.A.: Bayesian multitask inverse reinforcement leraning. arXiv:1106.3655v2 (2011)
7. Bellman, R.: A Markovian decision process. Journal of Mathematics and Mechanics **6**(5) (1957) 679–684
8. Puterman, M.L.: Markov Decision Processes. Wiley (1994)
9. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. 3rd edn. Pearson (2010)