Optimal Dynamic Treatment Regime by Reinforcement Learning in Clinical Medicine

David Han, Ph.D., Mina Song
The University of Texas at San Antonio, 1 UTSA Circle, San Antonio, TX 78249

Abstract
Precision medicine allows personalized treatment regime for patients with distinct clinical history and characteristics. Dynamic treatment regime implements a reinforcement learning algorithm to produce the optimal personalized treatment regime in clinical medicine. The reinforcement learning method is applicable when an agent takes action in response to the changing environment over time. Q-learning is one of the popular methods to develop the optimal dynamic treatment regime by fitting linear outcome models in a recursive fashion. Despite its ease of implementation and interpretation for domain experts, Q-learning has a certain limitation due to the risk of misspecification of the linear outcome model. Recently, more robust algorithms to the model misspecification have been developed. For example, the inverse probability weighted estimator overcomes the aforementioned problem by using a nonparametric model with different weights assigned to the observed outcomes for estimating the mean outcome. On the other hand, the augmented inverse probability weighted estimator combines information from both the propensity model and the mean outcome model. The current statistical methods for producing the optimal dynamic treatment regime however allow only a binary action space. In clinical practice, some combinations of treatment regime are required, giving rise to a multi-dimensional action space. This study develops and demonstrates a practical way to accommodate a multi-level action space, utilizing currently available computational methods for the practice of precision medicine.

Key Words: dynamic treatment regime, precision medicine, Q-learning algorithm, reinforcement learning

1. Reinforcement Learning

The reinforcement learning (RL) is a machine learning (ML) method that takes action in response to the changing environment over time for maximizing rewards, *R*; see Figure 1 below. The formulation of RL requires the policy defining a map from state to action, and the value function to calculate the total expected reward over time. The application domains of RL include

- **dynamic treatment regime** (**DTR**);
  - chronic diseases: cancer, diabetes, anemia, HIV, mental illness such as epilepsy, depression, Schizophrenia, opioid addiction
  - critical care: sepsis, anesthesia, ventilation, heparin dosing, and so on
- automated medical diagnosis with structured data (medical imaging) and unstructured data (free text);
- resource scheduling and task allocation, optimal process control, drug discovery (*de novo* design), healthcare management, etc.

2. Dynamic Treatment Regime

The dynamic treatment regime (DTR) is a RL approach in *precision medicine* to enable the optimal personalized treatment regime for patients with distinct genetic, demographic, and clinical characteristics.

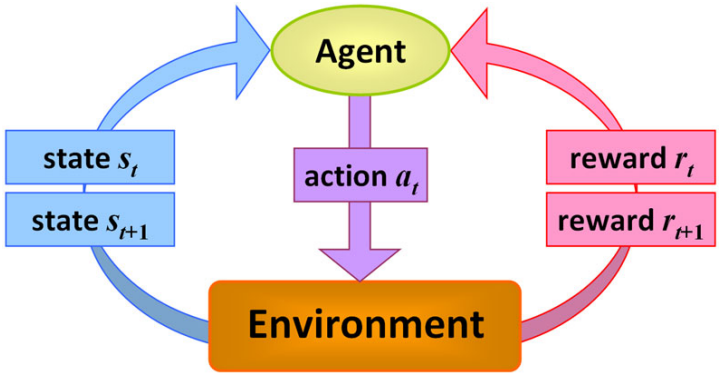

Figure 1: Schematic illustration of the reinforcement learning (RL)

3. Why RL-based DTR?

To produce the optimal personalized treatment regime in clinical medicine, RL is an effective approach for DTR due to several reasons.

- incomplete knowledge of the environment, which is usually estimated
  - The dynamic programming is often inappropriate.
- a limited sample size and costly data collection
- a causal association of historical conditions with the final outcome (*viz.*, no Markov property)
  - The state and action space grow exponentially, compared to the sample size.

4. Q-Learning

Q-learning is one of the popular methods to develop the optimal dynamic treatment regime (DTR). It is a temporal difference control algorithm to search for the optimal DTR based on longitudinal datasets. For estimation, it implements the backward recursive fitting of linear models based on a dynamic programming algorithm. It is mathematically formulated as

$$Q_t(h_t, a_t) = E\left[\max_{a_{t+1}} Q_{t+1}(h_{t+1}, a_{t+1}) \middle| h_t, a_t\right]$$

with $Q_T(h_T, a_T) = E[R|h_T, a_T]$. Since the algorithm is based on linear models, it provides easy implementation as well as easy interpretation for domain experts. To reduce the risk of model misspecification, several robust algorithms to the model misspecification have been developed. The inverse probability weighted estimator (IPWE) is one of such, and it estimates the mean outcome non-parametrically with different weights to the observed outcomes. It is robust but noisy contrast for classification. The other method is the augmented inverse probability weighted estimator (AIPWE), which combines the information from both propensity score and mean outcome models for smoothing. The fundamental limitation of these approaches is that the action space is *binary*, and it is strongly desired to implement a multi-dimensional action space for combinations of various treatment regimes.

5. Illustrative Example

Let us illustrate the proposed method using a two-stage treatment with multiple (3) treatments in each stage. With the sample size of n = 500, the following information is available for each stage; see Table 1 for the snapshot of a dataset.

stage 1: 3 covariates (x11, x12, x13)
3 treatments/actions (a11, a12, a13)
stage 2: 3 covariates (x21, x22, x23)
3 treatments/actions (a21, a22, a23)

Table 1. Snapshot of the dataset

x11	x12	x13	A1	x21	x22	x23	A2	R
-0.02	0.36	0.36	3	0.66	0.39	-0.38	2	1.61
0.49	-0.22	-0.28	2	-1.47	0.66	0.79	3	-0.41
0.75	0.55	-0.29	3	0.22	1.66	0.14	2	1.14
0.28	1.83	0.47	3	0.81	1.78	0.95	2	1.42
0.00	0.30	-0.09	2	0.30	0.33	1.00	1	0.31
				.	.			
				.	.			
				.	.			

The final outcome (reward) is *R*, which is a continuous variable we need to maximize. The higher the value is, the better the outcome is. The *empirical* treatment decision (action) is *A*, having a multinomial distribution with the probability vector given by

$$e^{X\beta_i} / \sum_j e^{X\beta_j}$$

The *optimal* treatment decision rule is *optA*, given for each stage as

stage 1: treatment 1 if (x11 > -0.54) and (x12 < 0.54)
else treatment 2 if (x11 > -0.54) and (x13 < 0.54)
else treatment 3
stage 2: treatment 1 if (x21 > 0.3) and (x23 < 0.46)
else treatment 2 if (x22 > 0.3) and (x23 < 0.46)
else treatment 3

Based on the proposed methodology, it was found out that the optimal DTR at stage 2 has a significantly better accuracy than the empirical treatment decision. The optimal DTR at stage 1 also has a significantly better accuracy than the empirical treatment decision.

Confusion matrix @ stage 2

		optA 1	2	3	
A	1	69	24	84	
	2	31	13	106	
	3	23	38	112	(accuracy = 0.388)

		optA 1	2	3	
DTR	1	108	2	28	
	2	3	46	18	
	3	12	27	256	(accuracy = 0.820)

Confusion matrix @ stage 1

		optA 1	2	3	
A	1	66	39	44	
	2	120	19	40	
	3	63	26	83	(accuracy = 0.336)

		optA 1	2	3	
DTR	1	241	67	64	
	2	8	0	9	
	3	0	17	94	(accuracy = 0.670)

References

Fernandez, K.C., Fisher, A.J., and Chi, C. (2017). Development and initial implementation of the dynamic assessment treatment algorithm. *PLoS One*, **12:** e0178806.

Laber, E.B. and Davidian, M. (2017). Dynamic treatment regimes, past, present, and future. *Statistical Methods in Medical Research*, **26:** 1605–1610.

Laber, E.B., Lizotte, D.J., Qian, M., Pelham, W.E., and Murphy, S. (2014). Dynamic treatment regimes: technical challenges and applications. *Electronic Journal of Statistics*, **8:** 1225–1272.

Murphy, S. (2003). Optimal dynamic treatment regimes. *Journal of Royal Statistical Society B*, **65:** 331–366.

Wallace, M.P. and Moodie, E.E. (2015). Doubly-robust dynamic treatment regimen estimation via weighted least squares. *Biometrics*, **71:** 636–644.

Zhang, Y., Laber, E.B., Davidian, M., and Tsiatis, A.A. (2018). Interpretable dynamic treatment regimes. *Journal of the American Statistical Association*, **113:** 1541–1549.

Zhang, Z. (2019). Reinforcement learning in clinical medicine: a method to optimize dynamic treatment regime over time. *Annals of Translational Medicine*, **7:** e345.