Reinforcement Learning in Observational Bio/Medical Applications

Elynn Chen

May 7, 2019

1 Introduction

The problem of treatment in medical condition is a online decision problem. For example, as illustrated in Figure 1, faced with a patient with sepsis, the physician must decide whether and when to initiate and adjust treatments such as antibiotics, intravenous fluids, vasopressor agents and mechanical ventilation.

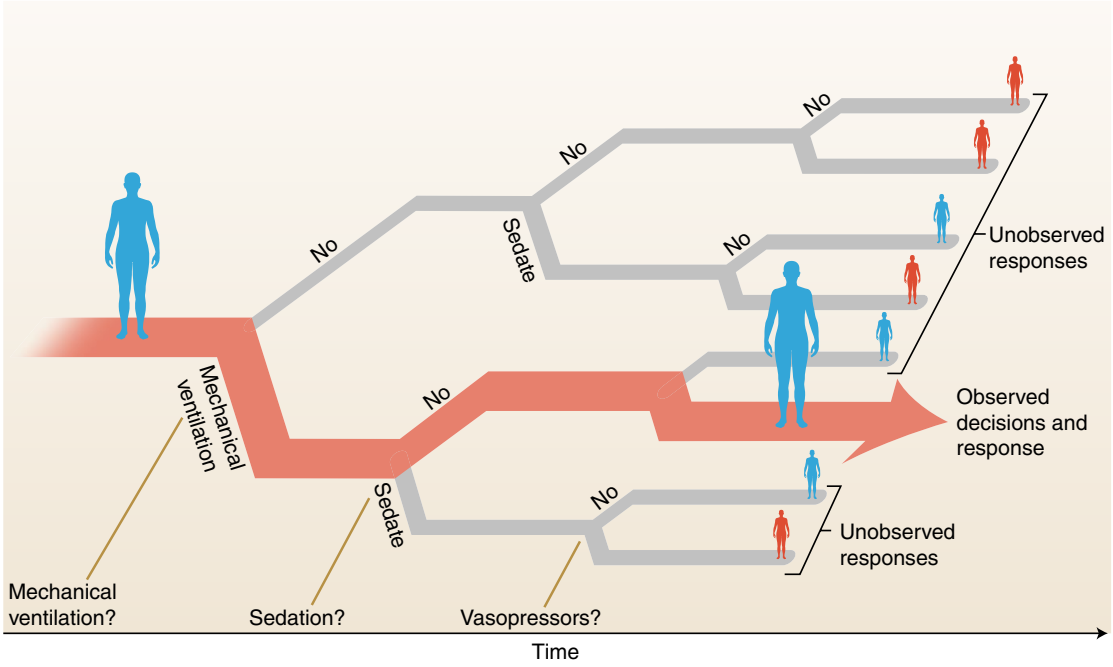


Figure 1: Sequential decision-making tasks. Image obtained from Gottesman et al. (2019). Credit: Debbie Maizels / Springer Nature

Sepsis is defined as severe infection leading to life-threatening acute organ dysfunction. The management of intravenous fluids and vasopressors in sepsis is key clinical challenge and a top re-

search priority. Currently, these decisions are made heuristically from physicians' years of trainings and experiences. Besides general guidelines, such as Surviving Sepsis Campaign, there is no tool existing currently to personalize treatment of sepsis and assist physicians in make decisions in real time and at the patient level. The widespread adoption of electronic health records (EHRs) opens new opportunity for a data-driven approach to health care, potentially creating new tools to aid physicians in deciding dynamic treatments.

Reinforcement learning (RL) provides tools to optimize sequences of decisions for long-term outcomes.

The aim of this project is to develop a decision support tool that leverages available patient information in the data-rich ICU setting to alert clinicians when a patient is ready for initiation of weaning, and to recommend a personalized treatment protocol. We build and validate a dynamic decision support system for health care based on the retrospective analysis of two nonoverlapping intensive care data bases containing data collected from adult patients.

2 Observational Bio/Medical Application

2.1 Intensive Care Unit Data

The Intensive Care Unit (ICU) data we use are the Medical Information Mart for Intensive Care version III (MIMIC-III) Database (Johnson et al., 2016) and the eICU Collaborative Research Database Pollard et al. (2018). MIMIC-III is a freely available source of de-identified critical care data from 53,423 adult admissions and 7,870 neonates from 2001 – 2012 in six ICUs at a Boston teaching hospital. The eICU Collaborative Research Database includes more than 3.3 million admissions from 2003 – 2016 in 459 ICUs across the United States. Both database contain high-resolution patient data, including demographics, time-stamped measurements from bedside monitoring of vital signs, laboratory tests, illness severity scores, medications and procedures, fluid intakes and outputs, clinician notes and diagnostic coding.

We use the the Medical Information Mart for Intensive Care version III (MIMIC-III) for model development, and the eICU Research Institute Database (eRI) for model testing, following the same data processing procedure in Komorowski et al. (2018). Specifically, the adult patients included in the analysis satisfy the international consensus sepsis-3 criterion. The training data extracted from MIMIC-III includes 17,083 unique ICU admissions from five separate ICUs in one tertiary teaching

hospital, while the testing data extracted from eICU contains 79,073 admissions from 128 different hospitals.

2.2 Problem Definition

In this paper, we seek to learn an dynamic treatment strategy by reinforcement learning for sepsis – a severe infection leading to life-threatening acute organ dysfunction. The management of intravenous fluids and vasopressors in sepsis is a key clinical challenge and a top research priority. So here we will also focus on the treatment choices of intravenous fluids (IV fluids) and vasopressors. The same problem was considered in Komorowski et al. (2018). Our definition of the sepsis cohort is the same as theirs. However, we employ different reward design and RL algorithm. Specifically, their methods estimates the transition matrix from training samples and use policy iteration and value iteration in standard dynamic programming, while ours also consider tabular Q-learning. Moreover, their definition of reward was only associated to the survive of a patient, while our reward design is more sophisticated. In Section 6, we outline future researches that follow the line of this work.

The sepsis patient cohort is selected according to the sepsis-3 criteria. We adopt the same definition and temporal criteria as described in Komorowski et al. (2018):

> Sepsis was defined as a suspected infection (prescription of antibiotics and sampling of bodily fluids for microbiological culture) combined with evidence of organ dysfunction, defined by a SOFA score ≥ 2. We adhere to the original temporal criteria for diagnosis of sepsis: when the antibiotics was given first, the microbiological sample must have been collected within 24h; when the microbiological sampling occurred first, the antibiotic must have been administrated within 72h. The earlier event defined the onset of sepsis. In line with previous research, we assumed a baseline SOFA of zero for all patient.

Each patient in the cohort is characterized by a set of 47 variables including demographics, Elixhauser premorbid status, vital signs, laboratory values, fluids and vasopressors received. Demographic information includes age, gender, weight. Vital signs include heart rate, systolic/diastolic blood pressure, respiratory rate et al. Laboratory values include glucose, total bilirubin, (partial) thromboplastin time et al. See variable.csv file for more information. For actions, we specifically consider the dosages of IV fluids and vasopressors.

All features were checked for outliers and errors using a frequency histogram method and univariate statistical approaches (Tukey's method). Errors and missing values are corrected when possible. For example, conversion of temperature from Fahrenheit to Celsius degrees and capping variables to clinically plausible values.

In the final processed data set, we have 17 621 unique ICU admissions, corresponding to unique trajectories fed into the RL algorithms.

3 The RL Approach to the Dynamic Treatment Decision

The dynamic treatment decision problem can be formulated as the problem of learning solution of an optimal control problem from a sample of trajectories. We consider the discrete-time optimal control problems for which the aim is to maximize a sum of discounted rewards over an infinite time horizon.

3.1 Markov Decision Process Formulation

The true patient physiological state is only partially represented by the data available, and therefore the disease process could be formulated as a partially observable Markov Decision Process (MDP). The MDP is characterized by the tuple $\{\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}\}$:

- **State space** $\mathcal{S}$, discrete or continuous. In our mode, at each time t, $s_t \in \mathcal{S}$ characterizes a patient's health state .
- An **action space** $\mathcal{A}$, discrete or continuous. At each time t, the agent takes action $a_t \in \mathcal{A}$ that changes the current state to the next state s_{t+1}. Here, we consider actions related to the dose prescribed of intravenous (IV) fluids and vasopressors.
- A **transition function** $P(s_{t+1}|s_t, a_t)$ is the probability of the next state s_{t+1} given the current state s_t and action taken a_t. It defines the (unknown) dynamics of the system.
- A **reward function** $R(s_t, a_t) \in \mathbb{R}$, the observed feedback received following a transition at each time step t. Transitions to desirable states yield a positive reward, and reaching undesirable states generates a penalty.

3.2 Generating $(s_t, a_t, s_{t+1}, r_{t+1}, done)$ Tuples from Observational Samples

In this section, we define the corresponding MDP elements in out ICU sepsis treatment application and explain the methods to extract them from the observational samples.

3.2.1 Irregular Observational Time Series Data

For each ICU admission, we code patient's data as multivariate discrete time series with 4 hours time step. Each trajectory covers from up to 24h preceding until 48h following the estimated onset of sepsis, in order to capture the early phase of its management, including initial resuscitation. The medical treatments of interest are the total volume of intravenous fluids and maximum dose of vasopressors administered over each 4 hour period. We use a time-limited parameter specific sample-and-hold approach to address the problem of missing or irregularly sampled data. The remaining missing data were interpolated in MIMIC-III using multivariate nearest-neighbor imputation. After processing, we have in total 278 598 sampled data points for the entire sepsis cohort.

3.2.2 State and Action Space Characterization

The state s_t is a 47-dimensional feature vector including fixed demographic information (age, weight, gender, admit type, ethnicity et al), vitals signs (heart rate, systolic/diastolic blood pressure, respiratory rate et al), and laboratory values (glucose, Creatinine, total bilirubin, partial thromboplastin time, paO_2, $paCO_2$ et al.).

The action a_t is a 2-dimensional vector, of which the first entry $a_t[0]$ specifies the dosages of IV fluids and the second $a_t[1]$ indicates the dosages of IV fluids and vasopressors, to be administrated over the next 4h interval.

In reality, most of the patient's measurements, the dosages of IV fluids and vasopressors are continuous variables. As a first step, we consider discrete state and action space in this paper.

The discrete state space is defined by clustering all patient time series from the training data by k-means method. We adopt 750 discrete mutually exclusive patient states. Two absorbing states were added to the state space, corresponding to the death and discharge of the patient. Thus, we have 752 states in total.

For action space, we discretize two variables into five actions respectively according to the Table **??**. The combination of the two drugs makes $5 \times 5 = 25$ possible actions in total.

Note that the above simplification of state and action space is only a first attempt. We expect

to expand them to allow continuous state and action space later. However, we may need more data samples.

3.2.3 Reward Design

The reward signal is important and need crafted carefully in real applications. Komorowski et al. (2018) uses hospital mortality or 90-day mortality as the sole defining factor for the penalty and reward. Specifically, when a patient survived, a positive reward was released at the end of each patient's trajectory (a reward of 100); while a negative reward (a penalty of -100) was issued if the patient died. However, this reward design is sparse and provide little information at each step. Also, mortality may correlated with respect to the health statues of a patient. So it is reasonable to associate reward to the health measurement of a patient after an action is taken.

In this application, we build our reward signal based on physiological stability. Specifically, in our design, physiological stability is measured by vitals and laboratory values v_t with desired ranges $[v_{\min}, v_{\max}]$. Important variables related to sepsis include heart rate (HR), systolic blood pressure (SysBP), mean blood pressure (MeanBP), diastolic blood pressure (DiaBP), respiratory rate (RR), peripheral capillary oxygen saturation (SpO2), arterial lactate, creatinine, total bilirubin, glucose, white blood cell count, platelets count, (partial) thromboplastin time (PTT), and International Normalized Ratio (INR). We encode a penalty for exceeding desired ranges at each time step by a truncated Sigmoid function, as well as a penalty for sharp changes in consecutive measurements.

$$r_{t+1} = \sum_v C_1 \left[\frac{1}{1+e^{-(v_t - v_{min})}} - \frac{1}{1+e^{-(v_t - v_{max})}} + 0.5 \right] - C_2 \left[\max\left(0, \frac{|v_{t+1} - v_t|}{v_t} - 0.2\right) \right],$$

Here, values v_t are the measurements of those vitals v believed to be indicative of physiological stability at time t, with desired ranges $[v_{min}, v_{max}]$. The penalty for exceeding these ranges at each time step is given by a truncated sigmoid function. The system also receives negative feedback when consecutive measurements see a sharp change.

Remark 1. There are definitely improvements in shaping the reward space. For example, in medical situation, the definition of the normal range of a variable sometime depends demographic characterization. Also, sharp changes in a favorable direction should be rewarded.

4 Off-Policy Learning and Evaluation

Due to the sample size constraint (278 598 training tuples), we focus on the tabular off-policy learning in this paper. In the future work section, we proposed several direction to deal with the problem of data sparsity in bio/medical applications.

4.1 Model-Based Policy Iteration

Komorowski et al. (2018) learned a theoretical optimal policy (which they call the 'AI policy') for the MDP using in-place policy iteration. We also implemented this method here. This direct, model based estimates of the policy value are known to reduce variance but add bias to the estimate.

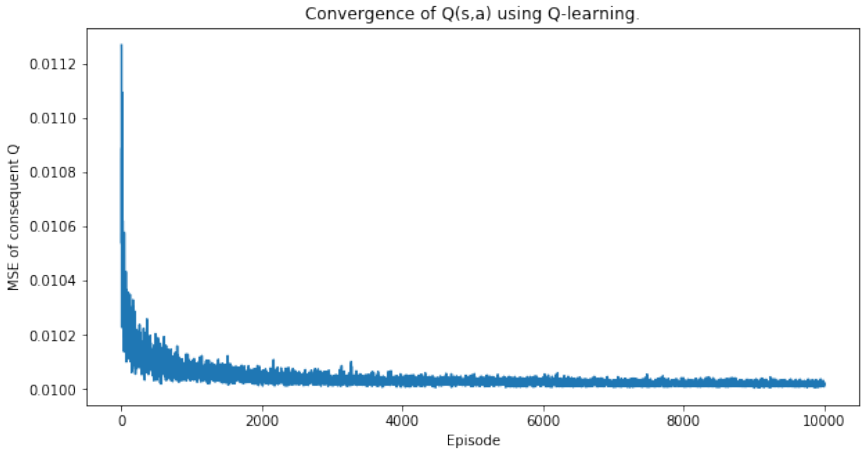
4.2 Tabular Q-learning

Specifically, we aim at estimating the expected value of state-action pairs $Q_\pi(s,a) : S \times A \to \mathbb{R}$. The best policy will be determined based on $\widehat{Q}_\pi(s,a)$. An off-policy algorithm starts with an initial state and arbitrary approximation of the Q-function, and update this estimate using the reward from the next transition using the Bellman update:

$$\widehat{Q}(s_t,a_t) = \widehat{Q}(s_t,a_t) + \alpha(r_{t+1} + \gamma \max_{a\in\mathcal{A}} \widehat{Q}(s_{t+1},a) - \widehat{Q}(s_t,a_t)),$$

where the learning rate α gives the relative weight of the current and previous estimate, and γ is the discount factor.

The following figure shows the convergence of Q-learning in experiments.



4.3 Off-Policy Evaluation

We evaluated the newly learned AI policy using a testing set of trajectories generated by physicians' policy. This is the off-policy evaluation problem. We implement the weighted importance

sampling (WIS) for off-policy evaluation and use bootstrapping to estimate the true distribution of the policy value in the test sets.

5 Results

We build 100 different models by randomly split the the entire MIMIC-III data set to 80% training and 20% testing subsets. The AI policies are estimated on the training set, using policy iteration and Q-learning, respectively. Then, the learned AI policies are evaluated on the testing set using WIS. The distribution of AI policy values are obtained by bootstrapping 75% of the testing set for 100 times.

6 Future Work

6.1 Data Sparsity

Medical observational data is sparse. In this article, we circumvent the problem by discretizing the continuous state and action spaces, which constrains us from utilizing more sophisticated deep RL methods. One possible solution to data sparsity is to interpolate observational data using Gaussian process and increase the sampling frequency. For example, sample the data every 10 minutes will produce approximately 3 million training data. This will provide enough samples for continuous state and action space.

6.2 Model-based RL

Model-based RL can also provide solution to data sparsity. By incorporating an environment model (simulator), we can generate infinite amount of training data. However, one should be cautious in specifying and model because the a model create bias thus the learned policy may not apply to real situation.

6.3 Reward Engineering

Application of AI techniques in bio/medical applications requires a great amount of domain knowledge. Both AI and medical communities will benefit from collaboration across disciplines. Reward can be designed to reflect the desired treatment guideline in certain situation.

6.4 Off-Policy Evaluation

The WIS is one of the most straight forward method to carry out off-policy evaluation. It is worth trying some newly developed method for off-policy evaluation.

References

Gottesman, O., F. Johansson, M. Komorowski, A. Faisal, D. Sontag, F. Doshi-Velez, and L. A. Celi (2019). Guidelines for reinforcement learning in healthcare. *Nature medicine 25*(1), 16–18.

Johnson, A. E., T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark (2016). Mimic-iii, a freely accessible critical care database. *Scientific data 3*, 160035.

Komorowski, M., L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine 24*(11), 1716.

Pollard, T. J., A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi (2018, September). The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data 5*, 180178.