CRRT Medical Treatment Prediction by Applying Reinforcement Learning Methods

Background of The Project

- CRRT: Continuous renal replacement therapy (CRRT), also known as continuous blood purification (CBP), is a new blood purification method.Continuous Renal Replacement Therapy (CRRT) is a treatment option for patients in need of dialysis or fluid removal. It is typically only utilized in the ICU setting and patients require this particular therapy because of their hemodynamic instability. CRRT is a much slower type of dialysis than regular HD, as it pulls fluid or cleans the blood continuously, 24 hours a day, rather than over a 2-4 hr treatment. Some facilities only use this treatment option in ICU patients with renal failure, even if they are hemodynamically stable. This type of therapy relies on the bedside nurse, who has special training in this technology and the equipment. It requires you to be aware of how the patient responds to the treatment both metabolically and hemodynamically at all times. CRRT includes continuous arteriovenous and venous hemofiltration (CAVH, CVVH), continuous arteriovenous and venous hemodialysis (CAVDH, CVVDH), continuous arteriovenous and venous hemodialysis (CAVHDF, CVVHDF) and other modes. CRRT, mechanical ventilation and extracorporeal membrane lung (ECMO) are the three major life support technologies for critically ill patients.
- Reinforcement Learning: RL is the science of decision making. It is about learning the optimal behavior in an environment to obtain maximum reward. This optimal behavior is learned through interactions with the environment and observations of how it responds, similar to children exploring the world around them and learning the actions that help them achieve a goal. In the absence of a supervisor, the learner must independently discover the sequence of actions that maximize the reward. This discovery process is akin to a trial-and-error search. The quality of actions is measured by not just the immediate reward they return, but also the delayed reward they might fetch. As it can learn the actions that result in eventual success in an unseen environment without the help of a supervisor, reinforcement learning is a very powerful algorithm.
- RL in Medical: The idea of reinforcement learning method is to take action in response to the changing environment. In clinical medicine, this idea can be used to assign optimal regime to patients with distinct characteristics. In the field of statistics, reinforcement learning has been widely investigated, aiming to identify an optimal dynamic treatment regime (DTR).

Data Preparation

- Uses MIMIC IV Data set
- Uses PostgreSQL. [?]
- Linking results of CRRT+Sepsis
- Current calculation of CRRT (9131, 14)
- In order to complete the missing data, we used the SAITS model for imputation. <SELF-ATTENTION-BASED IMPUTATION FOR TIME SERIES> Based on Transformer structure, the model manages to design weighted combination of two diagonally-masked self-attention (DMSA) blocks in order to make use of temporal dependencies and feature correlations between each time steps. It also introduces joint-optimization training approach that combines Observed Reconstruction Task (ORT) and Masked Imputation Task (MIT) to facilitate imputation.

Loss function The object to minimizing the loss L combines the multiple imputation losses of different learned representations. Generally speaking, each mean absolute error(MAE) loss is calculated between the artificially missing values and their respective imputations, which are defined as below.

$$\ell_{\mathrm{MAE}}(\text{ estimation, target, mask }) = \frac{\sum_d^D \sum_t^T |\,(\text{ estimation } - \text{ target }) \odot \text{ mask }_t^d}{\sum_d^D \sum_t^T \text{mask}_t^d}$$

The L represents a combination of two losses from the two tasks, ORT and MIT. MIT is utilized to predict artificially missing values so that imputation could be as accurate as possible, while ORT reconstructs observed values so that the model converges to the distribution of data.

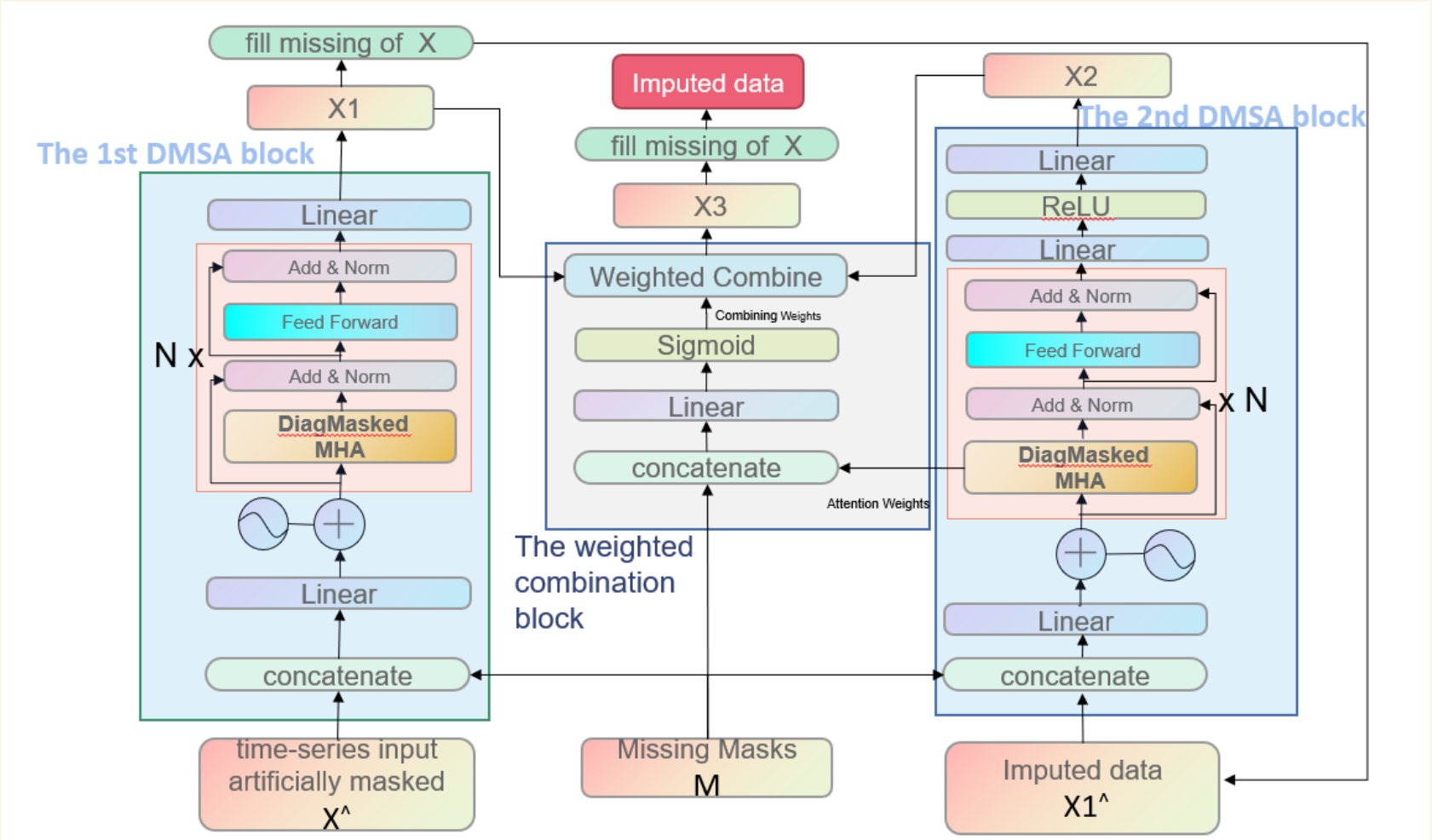
$$\mathcal{L} = \frac{1}{3}\left(\ell_{\mathrm{MAE}}\left(\tilde{X}_1, X, \hat{M}\right) + \ell_{\mathrm{MAE}}\left(\tilde{X}_2, X, \hat{M}\right) + \ell_{\mathrm{MAE}}\left(\tilde{X}_3, X, \hat{M}\right)\right) + \ell_{\mathrm{MAE}}\left(\hat{X}_c, X, I\right)$$

Dong Liu, Peter Zhao, Wenhan Yang, Yong Zhang, Ming Sheng, Huiying Zhao, Chenxiao Hao, Xu Yang

Beijing Institute of Technology, Tsinghua University, Peking University People's Hospital

CRRT Medical Treatment Prediction by Applying Reinforcement Learning Methods

- X is the original input with random missing values, while $\tilde{X}1$, $\tilde{X}2$, $\tilde{X}3$ are learned representations of different blocks of the model, and $\tilde{X}c$ is the final output about imputed data, which are described later. $\tilde{M}$ and I are the missing mask vectors containing 1 and 0, where 0 in $\tilde{M}$ used for ORT means missing for corresponding value and 1 in I used for MIT indicates artificially masked value.
  SAITS model utilizes the attention-mechanism to process data globally and parallelly, but the input of a single time step makes no contribution to its own values estimations. So diagonal masks are designed that diagonal attention weights approaching 0.
  SAITS adopts Positional Encoding and Feed-Forward Network like transformer, but it designs different blocks connected together as shown in the fig.



At first, the time-series vector artificially masked $X^{\wedge}$ and its missing mask vector M are concatenated as the input of the first DMSA block. The output, imputed data, is generated by missing values of $X^{\wedge}$ being filled with corresponding values in learned representation X1.
Besides to deepen the network to capture more correlations in time series, the learning target of the second DMSA block is also to verify these imputation values from the first DMSA block. But whether it can perform better than the first DMSA block or not is uncertain. So, to form the final presentation X3, the Weighted Combination Block combines X1 and X2 by weights that are processed from attention weights output by multi heads in the last layer of the second DMSA block. The final imputed data is generated in the same way as that in the first DMSA block.

Extraction of data sets

1. Access pressure: The pressure of blood before entering the blood pump. It mainly reflects the relationship between the blood flow provided by vascular access and the speed of blood pump. Arterial pressure is measured before the blood pump and measures the pressure (outside the body) when blood leaves the patient's blood access (such as a double-lumen catheter). Arterial pressure is measured to prevent excessive pumping of the blood pump and is typically -50 to -150 MMHG.
2. Effulent pressure: The pressure of blood outgoing the blood pump.
3. Ultrafiltrationrate (UFR) is the amount of solvent removed in plasma per unit time by ultrafiltration, expressed in ml/kg/h. At present, ultrafiltration rate is usually used to represent the therapeutic dose of CRRT. The number of milliliters of liquid passing through the membrane per hour.

Dong Liu, Peter Zhao, Wenhan Yang, Yong Zhang, Ming Sheng, Huiying Zhao, Chenxiao Hao, Xu Yang

Beijing Institute of Technology, Tsinghua University, Peking University People's Hospital

CRRT Medical Treatment Prediction by Applying Reinforcement Learning Methods

4. Citrate: Citrate is an important organic acid, colorless crystal, often contains a molecule of crystalline water, odorless, has a strong sour taste, easy to dissolve in water. The combination of citrate acid and 80℃ temperature has a good effect on killing bacterial spores, which can effectively kill the contaminated bacterial spores in the pipe of hemodialysis machine. Calcium ions must be involved in the formation of prothrombin activators and subsequent coagulation. Citrate ions and calcium ions can form a soluble complex which is difficult to dissociate, thus reducing the concentration of calcium ions in the blood and preventing blood coagulation.
5. Blood Flow: Blood flow is the rate at which blood is drawn from the body into the colander.
6. return_pressure : The Return Pressure of the fluid back into the body is an indicator of the patency of the venous return. It is usually positive, with a typical value of +50 to +150mmHg. Venous pressure is measured to prevent excessive resistance to blood return. The common causes of venous pressure alarm are: (1)the venous pipe is clamped or knotted; (2) Intraductal coagulation or catheter deviation or adherence to the wall in the vessel; (3) The patient is moving or being moved; (4) too fast blood flow; (5) The venous baroreceptors fail.
7. dialysate_rate : Dialysate is a physiological solution composed of inorganic ions and glucose in the body. Dialysate concentrations of sodium and chloride are usually physiological, while concentrations of magnesium and phosphorus are usually lower than physiological concentrations to allow for the removal of these substances during dialysis.
8. filter_pressure: Filter pressure is the pressure of fluid out of the body.
9. hourly_ patient_ fluid_removal: Blood flow out of patient's body within an hour.
10. prefilter_ replacement_rate: Prefilter replacement solution will deliver into the blood flow at set rate and blood will be diluted. The replacement"fluid volume"will then removed by the effluent pump. Prefilter administration has two effects: Dilutes the concentration of solute entering the filter. And decreases the efficiency of solute removal.
11. postfilter_ replacement_rate: The replacement"fluid volume"willbe removed by the effluent pump and blood will be concentrated. Postfilter replacement solution will deliver replacement solution to "replace"the removed"volume"and replenish lost electrolytes.
12. replacement_rate: The sum of prefilter replacement rate and postfilter replacement rate.

Model Preparation

- Model I: Conv + LSTM
- Model II: Regression + RL
  - UCB
    - Self-defined action state
  - Lasso
    - Self-defined class
- Input: The patient's vital signs showed initial UFR
- Output: UFR FF

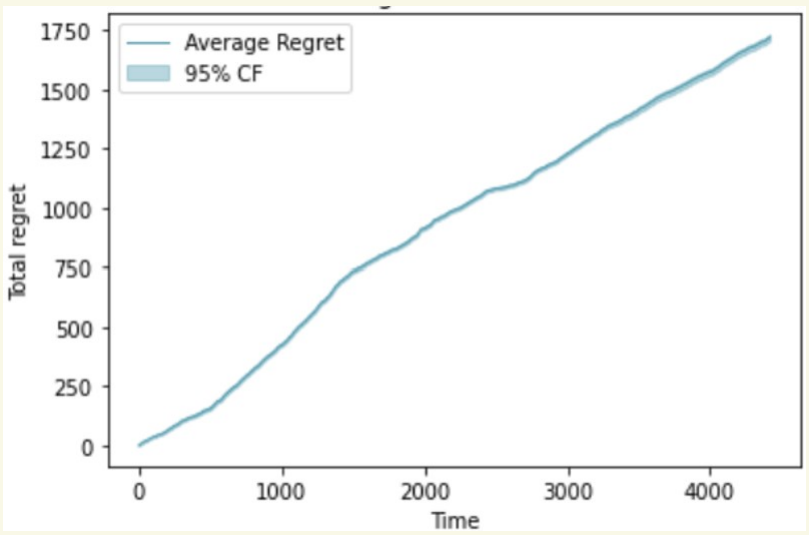


Figure: Ensemeble: regret as a function of time

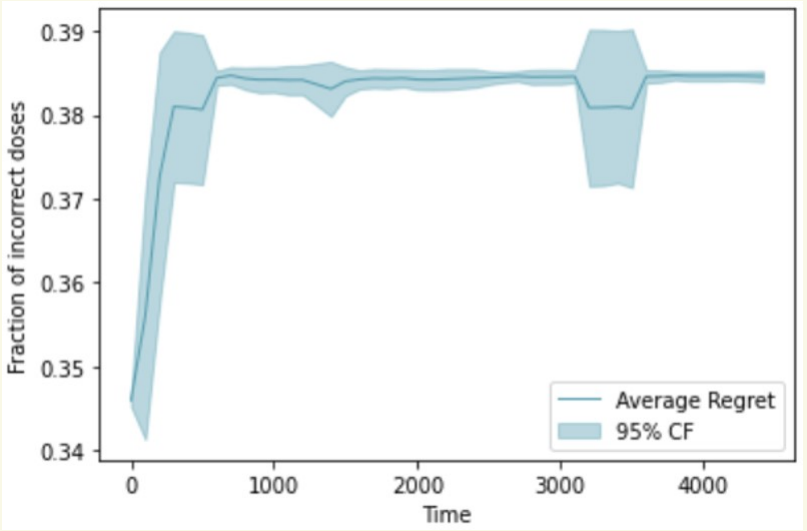


Figure: Fraction of incorrect does as a function of time

Dong Liu, Peter Zhao, Wenhan Yang, Yong Zhang, Ming Sheng, Huiying Zhao, Chenxiao Hao, Xu Yang

Beijing Institute of Technology, Tsinghua University, Peking University People's Hospital

Overleaf

CRRT Medical Treatment Prediction by Applying Reinforcement Learning Methods

The Problems to solve

- The most difficult task is dosage, here focus on dosage.
  - In this paper it has attempted to use reinforcement learning to give out a model that . The paper first examines two baselines: Conv + LSTM and a self-defined RL model. we implemented a Linear-UCB bandit that improved performance measured on regret and percent incorrect. Based on this model, we experimented with online supervised learning and reward reshaping to boost performance. The final results clearly beat the baselines and show promise of using multi-armed bandits and artificial intelligence to aid physicians in deciding proper dosages.

Input Output Of The Model

- Input
  - 'access_ pressure', 'blood_ flow', 'citrate', 'current_ goal', 'dialysate_ rate', 'effluent_ pressure', 'filter_ pressure', 'prefilter_ replacement_ rate', 'postfilter_ replacement_ rate', 'replacement_ rate', 'return_ pressure'
  - weights
- Output
  - 'hourly_ patient_ fluid_ removal'

Baseline

- Baseline I: Conv + LSTM
  - For a time-seris problem where we use multiple time series data to predict one time series columns, LSTM network with convolutional layers is applied to realize such prediction.
- Baseline II: CTRL
  - For the second baseline, we design a model called CTRL based on UCB algorithm and linear regression, besides, there are several techniques to be added into the model to enhance the performance of the prediction. Apart from the 13 CRRT treatment feature, the predetermined weights for each feature is also required as the input of the model.

LSTM

Long short-term memory (LSTM) is a special kind of RNN, which is mainly used to solve the problem of gradient disappearing and gradient explosion in the process of long sequence training. LSTMS perform better in longer sequences than ordinary RNN.
Whereas RNN has only one transfer state (h^t), LSTM has two states, one c^t(cell state) and one h^t(hidden state). The passed cell statec^t changes slowly. Usually, the output c^t is c^{t-1} passed from the previous state plus some values.
Hidden states h^t, on the other hand, often differ greatly under different nodes.
There are three main stages within LSTM:
-1. Forget phases. This phase is mainly about selective forgetting of the input passed by the previous node. Simply put, "forget what's not important and remember what's important." Specifically, we use the calculated z^f(f for forget) as a forget gate to control which c^{t-1} of the previous state should be kept and which should be forgotten.
-2. Memory choosing phase. This phase selectively "memorizes" the input of this phase. It's basically selective memory for the input x^t. Write down what is important, and write down less what is not. The current input is represented by the z calculated earlier. The selected gated signal is controlled by z^i (i stands for information). Add the results from the previous two steps to get the transfer to the next state.
-3. Output phase. This phase determines what will be considered the output of the current state. It's mainly controlled by z^O. The c^O obtained in the previous stage is also reduced. Similar to ordinary RNN, the output y^t is often eventually obtained by changing h^t.

Figure: Conv+LSTM model to realize CRRT prediction

Dong Liu, Peter Zhao, Wenhan Yang, Yong Zhang, Ming Sheng, Huiying Zhao, Chenxiao Hao, Xu Yang

Beijing Institute of Technology, Tsinghua University, Peking University People's Hospital

Overleaf

CRRT Medical Treatment Prediction by Applying Reinforcement Learning Methods

CTRL

- Consider the following learning problem: Make a choice among k actions repeatedly. After each choice is made, a certain value of revenue will be obtained, which is generated by the stationary probability distribution determined by the selected action. The goal is to maximize the expectation of total revenue over a certain period of time. This is the original form of the K-arm gambling machine problem. In the k-arm problem, each of the k actions has an expected or average payoff when selected, called the value of the action, and is denoted as: $q*(a)$.The action selected at the moment t is At and the corresponding payoff is Rt. At time t, the selected action is At, and the corresponding payoff is R_t. $q*(a) = E(a)[Rt|At = a]$
- If you know the value of each move, solving the K-arm gambling machine problem is simple: choose the move with the highest value each time.The actual problem can not exactly know the value of the action, can only be estimated. Let the estimate of the value of action a at time t be $Qt(a)$, and the closer the value is to $q*(a)$,the more accurate the estimate will be. If the value of actions is continuously estimated, there will be at least one action with the highest estimated value at any time, and these actions corresponding to the highest estimated value are called greedy actions. If greedy activities were selected, they could be called Exploitation, but if not, Exploration could be called.
  There are two parts of UCB value: $Qt(a)$+$Ut(a)$
  $Qt(a)$: Represents the actual distribution of the current action - return, thus the actual function Q.
  $U_t(a)$: A measure of uncertainty about the action.

$$At = \text{argmax}\, a|Qt(a) + U_t(a)|$$

$$= \text{argmax}\, a|Qt(a) + c\sqrt{\frac{lnt}{N_t(a)}}|$$

- c (c>0) is the control factor of exploration intensity, determines the confidence index.$Nt(a)$ is the number of times action a is selected at time t, it is easy to find when this value increase, $Ut(a)$ will decreases. If a is not chose, when t grows $U_t(a)$ will increase.
- Bad Action & State Group: For those treatment that does not make high-quality decisions, or did not make the best treatment decision in time, we classified them as bad action. We evaluate the treatment performance by comparing the vital sign indicators of the patients after the treatment, for each feature we classify the value into group "High", "Relative High", "Medium", "Relative Low", "Low", and for different group of vital sign indicators we assign different rewards to them, which will impose an influence the next step action computation to correct the former bad action.
- Delayed Reward: In medicine, the treatment often takes a period of time to reflect the curative effect, so the medical behavior cannot provide timely feedback on the treatment of the patient. From such perspective, we designed a delayed reward: in each step we pass a part of the reward to the next time step to measure long-term effects of medical behavior of the current treatment plan.

Dong Liu, Peter Zhao, Wenhan Yang, Yong Zhang, Ming Sheng, Huiying Zhao, Chenxiao Hao, Xu Yang

Beijing Institute of Technology, Tsinghua University, Peking University People's Hospital

Overleaf

CRRT Medical Treatment Prediction by Applying Reinforcement Learning Methods

Algorithm Implementation

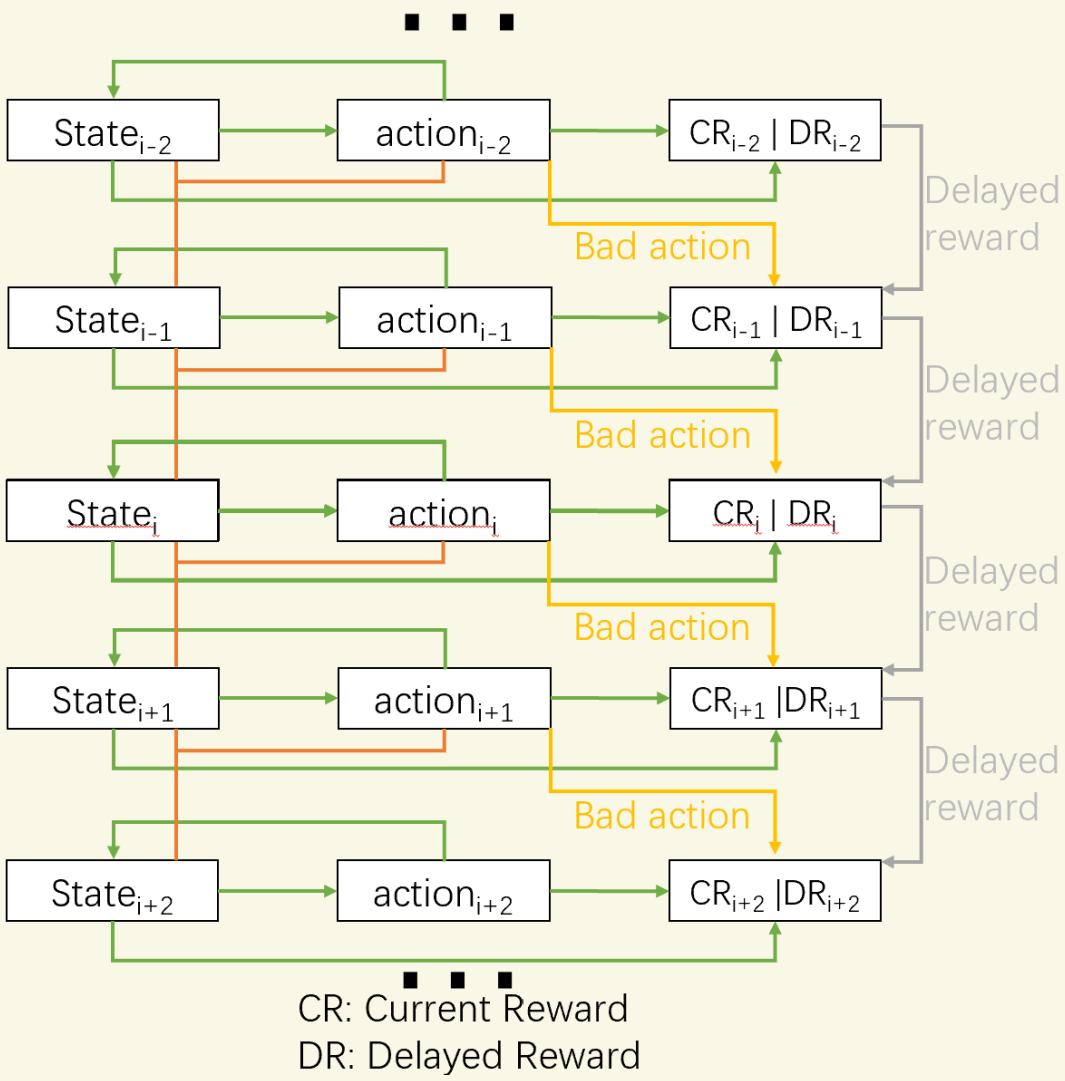


Figure: An example of CTRL

Algorithm 1 CTRL

Input: $a \in R_t$, Input CRRT features, reshape factor C

for $t = 1$ to n **do**
  observe context X_t_org
  $x_t \leftarrow group(X_t\_org)$
  $dr_t \leftarrow f(r_t - 1)$
  for each arm a **do**
    if a is new **then**
      $A_a \leftarrow Id$
      $b_a \leftarrow 0$
    end if
    $\hat{\theta} \leftarrow A_a^{-1} b_a$
    $p_{t,a} \leftarrow x_t^T \hat{\theta}_a + \sqrt{x_t^T A_a^{-1} x_t}$
  end for
  choose action $a_t \leftarrow \mathrm{argmax}_a p_{t,a}$
  observe reward r_t
  $calculate\Delta$ between expected action and a_t
  $r_t \leftarrow r_t - e^{\Delta}.C$
  if a_{t-1} is bad **then**
    $r_t \leftarrow g(r_t, a_{t-1})$
  end if
  $A_{at} \leftarrow A_{a,t} + \|x_t a_t\|^2$
  $b_{at} \leftarrow b_{at} + r_t a_t$
end for

Dong Liu, Peter Zhao, Wenhan Yang, Yong Zhang, Ming Sheng, Huiying Zhao, Chenxiao Hao, Xu Yang

Beijing Institute of Technology, Tsinghua University, Peking University People's Hospital

Overleaf

CRRT Medical Treatment Prediction by Applying Reinforcement Learning Methods

Elements and Results of the Experiment

	Prediction Accuracy
CTRL	81.6%
LSTM	60.3%
CTRL+Reward Reshape	84.7%
LSTM+Reward Reshape	69.3%
CTRL+Group Reward	83.5%
LSTM+Group Reward	76.9%
CTRL+Bad Action Correct	84.1%
LSTM+Bad Action Correct	79.4%
CTRL+Delay Reward	83.8%
LSTM+Delay Reward	80.1%
CTRL+Reward Reshape+Group Reward	85.2%
LSTM+Reward Reshape+Group Reward	81.3%
CTRL+Reward Reshape+Bad Action Correct	91.9%
LSTM+Reward Reshape+Bad Action Correct	85.5%
CTRL+Group Reward+Bad Action Correct	90.6%
LSTM+Group Reward+Bad Action Correct	84.2%
CTRL+Reward Reshape+Group Reward+Bad Action Correct	93.4%
LSTM+Reward Reshape+Group Reward+Bad Action Correct	86.7%

Conclusion and Future Work

In this work, we proposed a model to realize the prediction of proper CRRT medical treatment. From our experimental effect, the best model can give about 94.3% accurate timely treatment, which means that our model can give an effective treatment plan with a high probability, which is very meaningful for assisting doctors in making treatment decisions and reducing the cure time of CRRT patients.

In data extraction, our data extraction comes from MIMIC IV data set. Because there are many missing values in this data set, we use the SAITS model to impute data.

In the implementation of the CTRL model, we learned from the UCB model and the Linear Regression model. At the same time, we added some technologies, such as group evaluation, bad action reward, and delayed loss, to improve the performance of the model.

In the future, we are trying to design a reinforcement learning model based on DDPG to predict CRRT treatment.

Dong Liu, Peter Zhao, Wenhan Yang, Yong Zhang, Ming Sheng, Huiying Zhao, Chenxiao Hao, Xu Yang

Beijing Institute of Technology, Tsinghua University, Peking University People's Hospital

CRRT Medical Treatment Prediction by Applying Reinforcement Learning Methods

Appendix

- MIMIC dataset:

Medical Information Mart for Intensive Care is a large public database. MIMICS records data on patients in the intensive care unit of Beth Israel Female Deacon Medical Center from 2001 to 2019, has medical health data and records for more than 40,000 patients. The database records demographic information, such as a patient's gender, height, religion and other information. Recorded laboratory test information, such as blood routine, liver function, kidney function and other laboratory test data. The patient's medication information was recorded, such as hypertension patients taking hypertension drugs, etc. The level of care provided by the nursing staff and the patient is recorded.

- PostgreSQL Tool:

PostgreSQL is a free software object-relational database management system (ORDBMS) with complete features. It is based on POSTGRES 4.2 developed by the computer science Department of University of California.
PostgreSQL supports most SQL standards and offers many other modern features such as complex queries, foreign keys, triggers, views, transactional integrity, multi-version concurrency control, and more. Similarly, PostgreSQL can be extended in many ways, for example by adding new data types, functions, operators, aggregate functions, indexing methods, procedural languages, and so on.

Dong Liu, Peter Zhao, Wenhan Yang, Yong Zhang, Ming Sheng, Huiying Zhao, Chenxiao Hao, Xu Yang

Beijing Institute of Technology, Tsinghua University, Peking University People's Hospital

Overleaf