Yu *et al. BMC Medical Informatics and Decision Making* 2020, **20**(Suppl 3):124
https://doi.org/10.1186/s12911-020-1120-5

BMC Medical Informatics and Decision Making

RESEARCH Open Access

Supervised-actor-critic reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units

Chao Yu[1*†], Guoqi Ren[2†] and Yinzhao Dong[2]

From 5th China Health Information Processing Conference
Guangzhou, China. 22-24 November 2019

Abstract

Background: Reinforcement learning (RL) provides a promising technique to solve complex sequential decision making problems in healthcare domains. Recent years have seen a great progress of applying RL in addressing decision-making problems in Intensive Care Units (ICUs). However, since the goal of traditional RL algorithms is to maximize a long-term reward function, exploration in the learning process may have a fatal impact on the patient. As such, a short-term goal should also be considered to keep the patient stable during the treating process.

Methods: We use a Supervised-Actor-Critic (SAC) RL algorithm to address this problem by combining the long-term goal-oriented characteristics of RL with the short-term goal of supervised learning. We evaluate the differences between SAC and traditional Actor-Critic (AC) algorithms in addressing the decision making problems of ventilation and sedative dosing in ICUs.

Results: Results show that SAC is much more efficient than the traditional AC algorithm in terms of convergence rate and data utilization.

Conclusions: The SAC algorithm not only aims to cure patients in the long term, but also reduces the degree of deviation from the strategy applied by clinical doctors and thus improves the therapeutic effect.

Keywords: Reinforcement learning, Inverse learning, Mechanical ventilation, Sedative dosing, Intensive care units

Background

In the healthcare field, a clinical treatment plan consists of a series of decisions that determine the type of treatment and the dose of drug based on the current health condition and past treatment history of a patient. Therefore, the clinical treatment is usually characterized by a sequential decision-making process that lasts for a long period. RL aims to solve this kind of sequential decision-making problems when an agent chooses an action at each time step based on its current state, and receives an evaluative feedback and the new state from the environment [1]. In the past decades, applying RL for more efficient decision-making has become a hot research topic in healthcare domains [2], generating a great breakthrough in treatment of diabetics [3], cancer [4], sepsis [5], and many other diseases [6–8].

*Correspondence: yuchao3@mail.sysu.edu.cn
†Chao Yu and Guoqi Ren contributed equally to this work.
[1]School of Data and Computer Science, Sun Yat-Sen University, Guangzhou 510015, China
Full list of author information is available at the end of the article

© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

In the clinical treatment, clinicians should achieve the long-term goal of curing the patients, but at the same time, the dosage of the medicine should be controlled in daily range so as to maintain the stable condition of patients. The short-term object is also critical as it can avoid additional risk to patients due to improper dosage. Traditional RL, however, mainly considers inherent time delay that is assessed by long-term goals, but lacks of consideration of short-term effect. This may lead to a large cumulative reward for the learned strategy, but this strategy can deviate from the clinical treatment strategy significantly. In this paper, a Supervised-Actor-Critic (SAC) [9] algorithm is applied to solve the above problem. The agent of SAC takes curing a patient as the long-term goal, and the deviation degree of the treatment between the SAC agent and the clinician as the short-term goal. An expert strategy is defined as a Supervisor, which is used to guide the learning process to reduce the additional treatment risk to patients.

In the following sections, we first introduce some recent research progress of applying RL methods in ICUs. Then, we present the data preprocessing details and formalize the decision making process of mechanical ventilation and sedative dosing in ICUs. We then introduce the SAC algorithm in detail and discuss the results and analysis between SAC and AC. Finally, we conclude this paper with future works.

Related work

The development of artificial intelligence (AI) techniques and data processing methods enable optimal diagnose, treat and mortality prediction of patients in ICUs [10]. As one of the core AI technologies, RL has been widely applied in realizing intelligent decision-making in ICUs [2]. The authors [11–13] applied RL algorithms in addressing the administration of *intravenous* (IV) and maximum *vasopressor* (VP) in sepsis treatment. Padmanabhan et al. [14, 15] proposed an RL-based control strategy for ICU sedation regulation. Prasad et al. [16] applied fitted Q iteration with extremely randomized trees to determine the best weaning time of invasive mechanical ventilation. Utomo et al. [17] proposed a graphical model that was able to show transitions of patient health conditions and treatments for better explanability, and applied RL to generate a real-time treatment recommendation in ICUs. Nemati et al. [18] used deep RL methods to calculate optimal unfractionated Heparin from sub-optimal clinical ICU data. Yu et al. [19] used inverse RL to infer the reward functions when dealing with mechanical ventilation and sedative dosing in ICUs. Chang et al. [20] proposed a Q-learning method that jointly minimized the measurement cost and maximized predictive gain, by scheduling strategically-timed measurements in ICUs. Unlike all the existing studies that only consider the long term effects of treatment using RL methods, we also consider the short-term effects of treatment in terms of deviation from the doctor's clinical treatment expectations, in order to ensure safety during the learning process.

Preprocessing

Ventilation and sedation dosing in ICUs

Effective ventilation is one of the most commonly used methods in the treatment of patients in ICUs. These patients are usually featured with acute respiratory failure or impaired lung function caused by some underlying factors, such as pneumonia, sepsis or heart disease. In addition, respiratory support is required after major surgery for consciousness disorders or weakness. Whether a patient is ready for extubation is determined by some major diagnostic tests, involving screening for potential disease resolution, hemodynamic stability, current ventilator assessment settings and awareness levels, and the final series of spontaneous breathing tests (SBTs). Serious discomfort and longer time stay in ICUs will occur if a patient must be reintubated due to the failure of breath test and other reasons within her first stay of 48 to 72 h.

In ICUs, another major treatment means is the use of sedative doses, which is essential for maintaining the patient's physiological stability. According to a relevant research, there is a certain correlation between the timing of ventilation and the application of sedatives in ICUs. Therefore, it is necessary to propose more effective ventilation and dosing methods so as to improve the patient's treatment effect, and reduce the patient's residence time and associated cost in ICUs.

Data processing

Firstly, we extract 8860 admissions from adult patients in MIMIC-III database [21], and exclude those admissions who were kept under ventilation for less than 24 hours, or failed being discharged from ICUs at the end of admission. The MIMIC is a free resource-rich ICU research database, which was first published in 2006 by the Computational Physiology Laboratory of the Massachusetts Institute of Technology (MIT), the Beth Israel Dikang Medical Center (BIDMC) and Philips Medical Center. It contains medical data of nearly 40,000 adults and 8,000 newborns in ICUs. The median age of adult patients was 65.8 years, of which 55.9% were males and 11.5% were hospitalized. The database is mainly used for academic and industrial research, offering a variety forms of data in ICU including demographic characteristics, vital signs, experimental testing, diagnosis, dosage of drugs, length of stay and other critical care unit data. We use support vector machines (SVM) [22] to fit the physiological measured values

at different measurement times. After preprocessing, we take 10 minutes as the frequency of time series from admission time to discharge time. Please refer to [19] for more details in data processing.

Formulation of the MDP

Following previous studies, the decision-making problem is modeled as an MDP by a tuple of $< S, A, P, R >$, where $s_t \in S$ is a patient's state at time t, $a_t \in A$ is the action made by clinicians at time t, $P(s_{t+1}|s_t, a_t)$ is the probability of the next state after given the current state and action, and $r(s_t, a_t) \in R$ is the observed reward following a transition at time step t. The goal of an RL agent is to learn a policy to maximize the expected accumulated reward over time horizon T by:

$$R^{\pi}(s_t) = \lim_{T\to\infty} E_{s_{t+1}|s_t,\pi(s_t)} \sum_{t+1}^{T} \gamma^t r(s_t, a_t)$$

where the discount factor γ determines the relative weight of immediate and long-term rewards.

The MDP of ventilation and sedation dosing in ICUs can be express as follow.

State: A patient's state is composed of 13-dimensional features, including respiration rate, heart rate, arterial pH, positive end-expiratory pressure (PEEP) set, oxygen saturation pulse oxymetry (SpO2), inspired oxygen fraction (FiO2), arterial oxygen partial pressure, plateau pressure, average airway pressure, mean non-invasive blood pressure, body weight (kg) and age.

Action: The two discrete actions regarding ventilation are defined as whether weaning off a patient from the ventilator. As for the sedative, the propofol was discretized into four different actions. Ultimately, there are eight action combinations.

Reward: The reward function r_{t+1} is defined as $r_{t+1} = r_{t+1}^{vitals} + r_{t+1}^{vent\ off} + r_{t+1}^{vent\ on}$ [16, 19], in which r_{t+1}^{vitals} evaluates the effect of these actions on the physiological stability of the patient within a reasonable range, $r_{t+1}^{vent\ off}$ estimates the performance of ventilation being stopped at time $t+1$, and $r_{t+1}^{vent\ on}$ simply represents the cost per hour on the ventilator.

Methods

Machine learning can be divided into three categories: supervised learning, unsupervised learning and RL. Supervised learning continuously reduces the error between the predicted value and the original value by the tagged data. The common problems of supervised learning application are classification and regression. Unsupervised learning, however, aims at finding correlation of data without labels in clustering and dimension reduction. Unlike traditional supervised learning methods that usually rely on one-shot, exhaustive and supervised reward signals, RL tackles with sequential decision making problems with sampled, evaluative and delayed feedback simultaneously [2]. The sequential decision making process of medical problems usually includes multiple steps in sequence, and RL is good at dealing with such problems. We can build an MDP model and use an RL algorithm to learn an optimal treatment strategy. The long-term goal of RL is to maximize the cumulative reward value, which means that patients must recover, but there is also a risk of drug use that deviates from clinician guidance. Therefore, we incorporate clinician guidance in the framework of RL such that the action of drug selection is in line with the guidance of clinicians.

Algorithm principle

The framework of SAC algorithm is shown in Fig. 1. Based on AC, a supervised learning mode (Supervisor) is added to the Actor network, which changes the gradient direction and updates the hyperparameters of the Actor network [23]. During the Actor network optimization, the Supervisor is optimized at the same time. The Critic computes the value functions based on the Reward and State in current Environment, then passes the TD error to the Actor. The Actor updates the strategy based on the Critic's TD error and the supervision error from the Supervisor.

Actor network update

The Actor network obtains the best strategy by updating hyper-parameters θ. The input of this network is state and the output is action. We use the TD error and supervised learning error to optimize hyper-parameter θ. To this end, we propose the following formula:

$$J(\theta) = (1-\epsilon)J_{RL}(\theta) + \epsilon\,(-J_{SL}(\theta))$$

where $J_{RL}(\theta)$ represents the optimization goal of an RL algorithm, which represents the reward value expectation of the trajectory under the current strategy, and $J_{SL}(\theta)$ is the optimization goal of supervised learning, which represents the degree of difference between the predicted action and the labeled action. $J_{SL}(\theta)$ is usually expressed in the form of variance or conditional entropy between the predicted value and the original value. ϵ is a weighting parameter to balance the contribution between RL and supervised learning.

Our goal is to maximize the reward value and make less difference in the actions between the RL agent's actions and the clinician's actions in the process of optimizing the strategy. We use the method of stochastic gradient descent [24] to optimize the parameter θ as follows:

$$\theta = \theta + \alpha\left((1-\epsilon)\frac{\partial J_{BL}(\theta)}{\partial\theta} + \epsilon\left(-\frac{\partial J_{SL}(\theta)}{\partial\theta}\right)\right)$$

where α represents the learning rate.

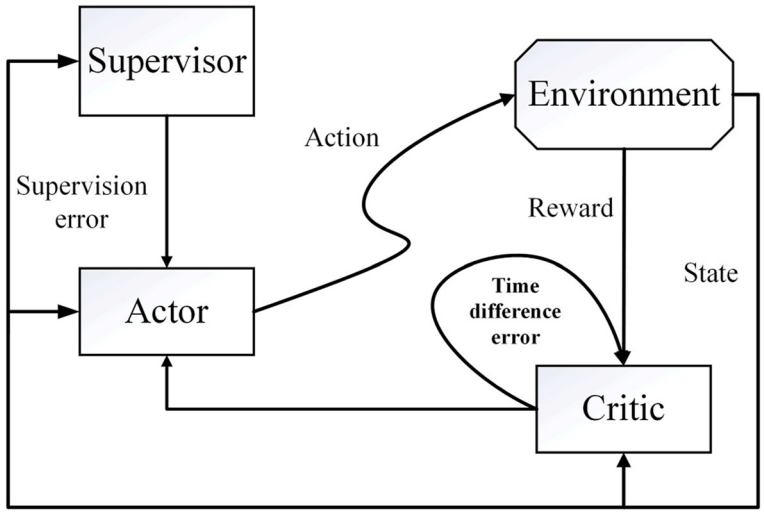

Fig. 1 The framework of SAC algorithm

Critic network update

The Critic network guides the learning of the Actor network, while the Actor network outputs the final treatment strategy. The Critic network estimates the action-state reward value $Q_w(s,a)$. During learning, the Critic network outputs a predicted Q value $Q(s_t,a_t)$ through $Q_w(s,a)$. The update of the Critic network parameter θ is as follows:

$$J(w) = E_{r_t,s_t}\left[\left(Q_w(s_t,a_t) - y^t\right)^2\right]$$

$$y_t = r\left(s_t,a_t\right) + \gamma Q_w^{tar}\left(s_{t+1},\mu_\theta(s_{t+1})\right)$$

where $J(w)$ is the loss function of the Critic network, and Q_w^{tar} is the target network parameter of the Critic network.

Results

Experimental setup

Since the feature is not always continuous and it may be a classification value, it is meaningless to compare such values. For example, [Red, Yellow, Blue] can be mapped to [0, 1, 2] to reflect their relationships but this mapping does not capture the relationship within the original feature attributes. This problem can be solved by using a one-hot encoding by using an N-bit status register to encode the N states, each state having its own separate register bit, and only one bit is active at any time [25]. Taking the above problem as an example, after using the one-hot encoding, [red, yellow, blue] can be encoded as [[0,0,1],[0,1,0],[1,0,0]]

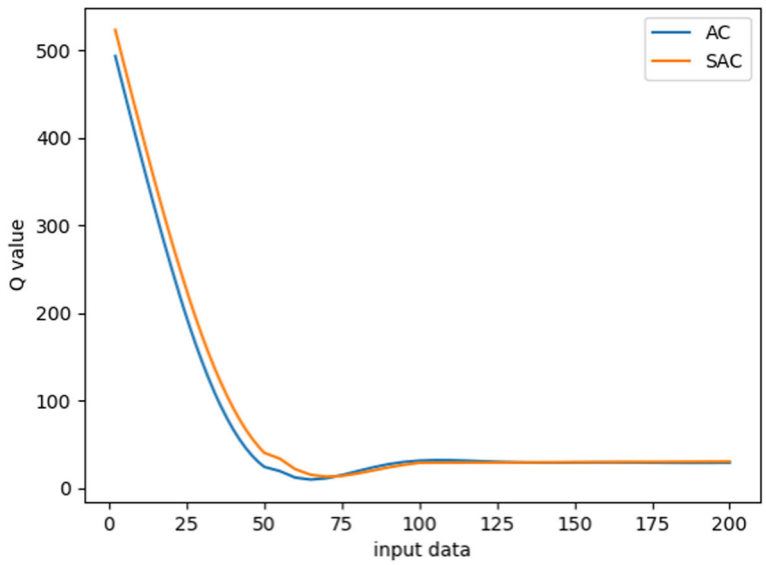

Fig. 2 Learning dynamics in terms of Q values regarding ventilation using SAC and AC algorithms

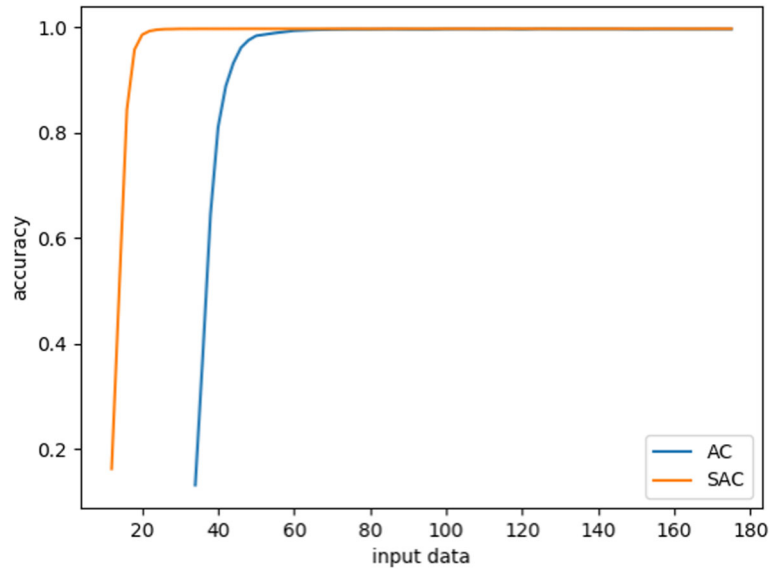

Fig. 3 The process of SAC and AC algorithms evaluating the accuracy of the test set on ventilation

so that the original relationship of features can be maintained. In our formulation, the actions for the ventilation and sedative doses are encoded using the one-hot encoding, separately.

In the MDP of ventilation, the state is the 13-dimensional patient's physiological characteristics, ventilator parameters, and current ventilation status. In AC and SAC, the Actor network, the Critic network and the Supervisor network all adopt a three-layer neural network. The Actor network has 20 neurons in the hidden layer with the ReLU activation function, and two neurons in the output layer with Softmax as the activation function. The Critic network has 20 neurons in the hidden layer, using the ReLU activation function, and a neuron in the output layer without using an activation function. The Supervisor network has 9 neurons in the hidden layer and two neurons in the output layer, and the Softmax is also used as the activation function for the output layer.

In the MDP model of sedative dose, the state is a 14-dimensional feature of the previous 13-dimensional feature combined with the ventilation action. The action is a sedative dose, which is encoded by the One-Hot encoding to form a four-category option. The network structure of it is the same as the MDP model of ventilation.

Experimental results and analysis

Figure 2 shows the results in the ventilation experiment, where the vertical axis is the Q value of the sample data, and the horizontal axis represents the training process. As the training proceeds, the Q value decreases gradually. Finally, both SAC and AC converge and reach a stable level. The trend of decline is basically the same, illustrating that both SAC and AC have a similar network structure for Q value prediction. However, SAC converges a bit faster than AC from a slightly higher initial values. Figure 3 shows that the Accuracy rate (AR) of the two algorithms has increased significantly with the increasing of episodes, and the stability level is above 95%. However, it can be seen that the convergence speed of the SAC algorithm is much faster than the AC algorithm. It takes 20 episodes for SAC to converge to 99% AR, and 60 episodes for AC. This is due to the Supervisor network in SAC, which can update the Actor network with simultaneous guidance of both the Supervisor part and the Critic part.

To further validate the effectiveness of SAC, we test the learned policy on the testing sets of expert and non-expert data sets. As shown in Table 1, both SAC and AC algorithms have an AR of over 99% on the testing set. The AR of SAC is slightly higher than that of AC, which indicates that the strategy learned by SAC is closer to the real medical strategy than that by AC. Meanwhile, it is worth noting that SAC is more accurate on expert data sets than on other data sets, while the AR of AC algorithm is comparatively balanced on each data set. It shows that the strategy learned by SAC is closer to the strategic plan of the experts.

Table 1 The AR of learned strategies using SAC and AC algorithms on the test data set

Strategy	Validation set	Expert data	Common single intubation	Multiple intubation
SAC	99.55%	99.57%	99.51%	99.55%
AC	99.48%	99.47%	99.46%	99.49%

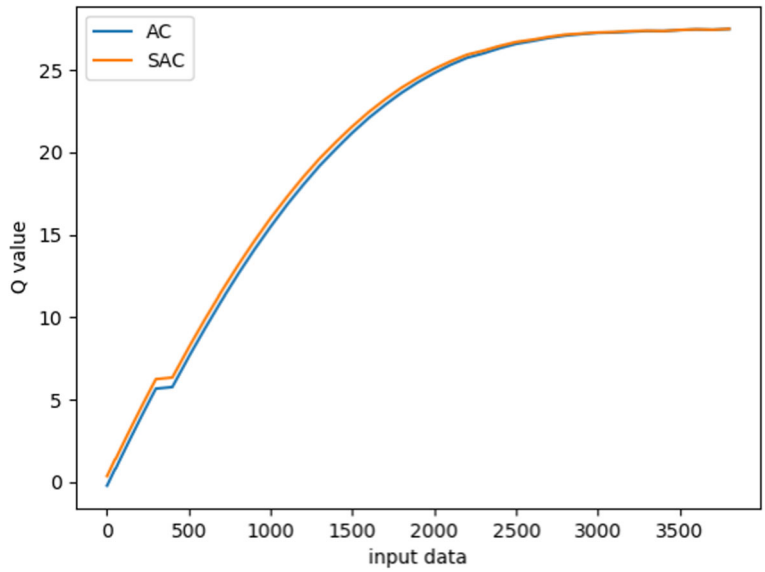

Fig. 4 Learning dynamics in terms of Q values regarding sedative using SAC and AC algorithms

Figure 4 gives the experiment results for the sedative dosing. Since the learning data and the network structure are consistent during the process of ventilation experiment, the convergence trend of the Q value of SAC and AC algorithms is roughly the same. Figure 5 shows the mean square error (MSE) between the predicted and original values of the sedative dose on the train set. It can be seen that both SAC and AC can converge to a stable MSE after 10000 episodes. However, after that, the convergence value of SAC is lower than that of AC, which indicates that the strategy learned by SAC is closer to the clinician's treatment strategy than that by AC. Besides, SAC is relatively more stable than AC in the final episode. This is because SAC introduces a supervised learning process such that a higher deviation from the clinician's strategy can be reduced to a smaller value. The AC algorithm without supervision network only aims at maximizing the cumulative reward value, causing great fluctuations in the learning process.

Discussions

We further analyze the differences between MSE and AR of two algorithms on the test set as shown in Table 2. In terms of AR, the SAC algorithm is slightly better than

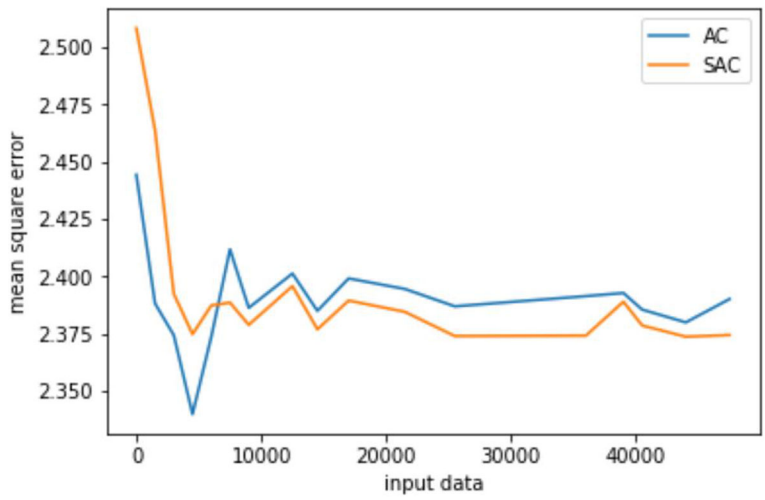

Fig. 5 MSE reduction process of SAC and AC algorithms on the train set for sedative dosing

Table 2 The AR and MSE of learned polices using SAC and AC algorithms on the test data set

Strategy	MSE	AR
SAC	2.49	41.5%
AC	3.10	41.5%

the AC algorithm, indicating that the strategy learned by SAC is closer to the real medical strategy than that of the AC algorithm. Although SAC and AC have the same AR on test set, SAC has a smaller MSE than AC. Thus, the SAC algorithm is closer to the real treatment strategy than the AC algorithm under the same AR. Table 3 shows the performance of two algorithms in expert data, single intubation and multiple intubation data. The performance of SAC is better than AC in MSE and AR on expert data set and non-expert data set. Especially, in multiple intubation data set, the AR of SAC is 6% higher than AC, and the MSE is reduced by 0.8. This illustrates that the goal of the clinician is indeed to cure the patient, but it is necessary to maintain a stable state of the patient under complex medical environments. Therefore, the introduction of supervised RL is more in line with the medical settings than the simple RL alone.

Conclusions

In this paper, we first introduce the principle and advantages of incorporating supervised learning into RL, and then establish the MDP for mechanical ventilation and the dose of sedatives for patients in ICUs. During the process of learning the strategies, SAC not only achieve the long-term goal of curing patients, but also meet the short-term goal of approaching the clinician's strategy gradually. Compared with the AC algorithm, SAC is more suitable to solve the problem of ventilation and the dose of sedatives in ICUs. Finally, we validate that SAC algorithm is slightly better than the AC algorithm in matching the clinician strategy, and its convergence speed and data utilization efficiency are much higher than AC. In the future, we will apply the SAC algorithm to other healthcare domains such as HIV and Sepsis to achieve more efficient and stable dynamic treatment regimes.

Table 3 The AR and MSE of learned polices using SAC and AC algorithms on expert data, single intubation and multiple intubation

Strategy	Expert data		Single intubation		Multiple intubation	
	MSE	AR	MSE	AR	MSE	AR
SAC	2.52	37%	2.71	38%	2.28	47%
AC	3.01	34%	3.15	35%	3.08	41%

Abbreviations

RL: Reinforcement learning; ICUs: Intensive care units; HIV: Human immunodeficiency viruses; SBTs: Spontaneous breathing tests; MIMIC: Medical information mart for intensive care; MIT: Massachusetts Institute of Technology; BIDMC: Beth Israel Dikang Medical Center; SVM: Support vector machine; MDP: Markov decision process; AC: Actor-critic; SAC: Supervised actor-critic; AR: Accuracy rate; MSE: Mean square error

Acknowledgements

Not applicable.

About this supplement

This article has been published as part of *BMC Medical Informatics and Decision Making Volume 20 Supplement 3, 2020: Health Information Processing*. The full contents of the supplement are available online at https://bmcmedinformdecismak.biomedcentral.com/articles/supplements/volume-20-supplement-3.

Authors' contributions

YC lead the research and participated in the manuscript revision. RG and DY carried out the method application, experiment conduction and result analysis. DY participated in the data extraction and preprocessing. All authors contributed to the preparation, review, and approval of the final manuscript and the decision to submit the manuscript for publication.

Funding

This work is supported by the Hongkong Scholar Program under Grant No. XJ2017028. The funding supports the authors' visit to Hongkong Baptist University, where the research was conducted, and covers the publication cost of this article.

Availability of data and materials

The datasets used and/or analysed during the current study is available from the first author on reasonable request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School of Data and Computer Science, Sun Yat-Sen University, Guangzhou 510015, China. [2]School of Computer Science and Technology, Dalian University of Technology, Dalian 110621, China.

Published: 9 July 2020

References

1. Sutton RS, Barto AG. Reinforcement Learning: An Introduction. Cambridge, Massachusetts: The MIT press; 1998.
2. Yu C, Liu J, Nemati S. Reinforcement learning in healthcare: A survey. 2019. arXiv preprint arXiv:1908.08796.
3. Bothe MK, Dickens L, Reichel K, Tellmann A, Ellger B, Westphal M, Faisal AA. The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. Expert Rev Med Devices. 2013;10(5):661–73.
4. Tseng HH, Luo Y, Cui S, Chien JT, Ten Haken RK, El Naqa I. Deep reinforcement learning for automated radiation adaptation in lung cancer. Med Phys. 2017;44(12):6690–705.
5. Yu C, Ren G, Liu J. Deep Inverse Reinforcement Learning for Sepsis Treatment. In: 2019 IEEE International Conference on Healthcare Informatics (ICHI). New York: IEEE; 2019. p. 1–3.
6. Shortreed SM, Laber E, Lizotte DJ, Stroup TS, Pineau J, Murphy SA. Informing sequential clinical decision-making through reinforcement learning: an empirical study. Mach Learn. 2011;84(1-2):109–36.
7. Nagaraj V, Lamperski A, Netoff TI. Seizure control in a computational model using a reinforcement learning stimulation paradigm. Int J Neural Syst. 2017;27(07):1750012.

8. Yu C, Dong Y, Liu J, Ren G. Incorporating causal factors into reinforcement learning for dynamic treatment regimes in HIV. BMC Med Inform Decis Making. 2019;19(2):60.
9. Konda VR, Tsitsiklis JN. Actor-critic algorithms. In: Advances in neural information processing systems. Cambridge: MIT Press; 2000. p. 1008–14.
10. Johnson AE, Ghassemi MM, Nemati S, Niehaus KE, Clifton DA, Clifford GD. Machine learning and decision support in critical care. Proc IEEE Inst Electr Electron Eng. 2016;104(2):444–66.
11. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med. 2018;24(11):1716–20.
12. Raghu A, Komorowski M, Ahmed I, Celi L, Szolovits P, Ghassemi M. Deep reinforcement learning for sepsis treatment. 2017. arXiv preprint arXiv:1711.09602.
13. Raghu A, Komorowski M, Celi LA, Szolovits P, Ghassemi M. Continuous State-Space Models for Optimal Sepsis Treatment: A Deep Reinforcement Learning Approach. In: Machine Learning for Healthcare Conference. Cambridge: MIT Press; 2017. p. 147–63.
14. Padmanabhan R, Meskin N, Haddad WM. Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning. Biomed Signal Process Control. 2015;22:54–64.
15. Padmanabhan R, Meskin N, Haddad WM. Optimal adaptive control of drug dosing using integral reinforcement learning. Math Biosci. 2019;309: 131–42.
16. Prasad N, Cheng LF, Chivers C, Draugelis M, Engelhardt BE. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. 2017. arXiv preprint arXiv:1704.06300.
17. Utomo CP, Li X, Chen W. Treatment Recommendation in Critical Care: A Scalable and Interpretable Approach in Partially Observable Health States. In: 39th International Conference on Information Systems. New York: Curran Associates; 2018. p. 1–9.
18. Nemati S, Ghassemi MM, Clifford GD. Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). New York: IEEE; 2016. p. 2978–81.
19. Yu C, Liu J, Zhao H. Inverse reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. BMC Med Inform Decis Making. 2019;19(2):57.
20. Chang CH, Mai M, Goldenberg A. Dynamic Measurement Scheduling for Event Forecasting using Deep RL. 2019. arXiv preprint arXiv:1901.09699.
21. Johnson AE, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035.
22. Shawe-Taylor J, Cristianini N. Support vector machines. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. United Kingdom: Cambridge university press; 2000, pp. 93–112.
23. Si J, Barto AG, Powell WB, Wunsch D. Supervised actor-critic reinforcement learning. In: Handbook of learning and approximate dynamic programming. London: IEEE Press; 2004. p. 359–80.
24. Zinkevich M, Weimer M, Li L, Smola AJ. Parallelized stochastic gradient descent. In: Advances in neural information processing system. Cambridge: MIT Press; 2010. p. 2595–603.
25. Golson S. One-hot state machine design for FPGAs. In: Proc. 3rd Annual PLD Design Conference & Exhibit, vol. 1. New York: IEEE; 1993.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions