HeartPole: A transparent task for reinforcement learning in Healthcare

Vadim Liventsev[1,2], Alexandre Simon [2], Aki Härmä [2] and Milan Petković, [1,2]
[1]*Eindhoven University of Technology, Eindhoven, the Netherlands*
[2]*Philips Research, Eindhoven, the Netherlands*
{*v.liventsev, m.petkovic*}*@tue.nl,* {*vadim.liventsev, alexandre.simon, aki.harma, milan.petkovic*}*@philips.com*

Keywords: Reinforcement learning, neural networks, clinical decision support, patient simulators

1 INTRODUCTION

Reinforcement learning in Healthcare is an emergent field that has created a demand for patient simulators like GYMIC (Kiani et al.,) - a black box neural model trained on MIMIC III dataset (Johnson et al., 2016) that predicts health outcomes of clinical decisions and can be used for training clinical decision-making models. We introduce a patient simulator inspired by *CartPole* (Barto et al., 1983) that trades clinical accuracy off for *simplicity* and *transparency*, while still being *non-trivial* to solve.

2 HEARTPOLE ENVIRONMENT

HeartPole simulates a creative professional trying to become more productive. However, many decisions that would help in the short term (not sleeping, consuming coffee and alcohol) can create long-term health issues that negate all short term gains.

HeartPole is a fully observable Markov Decision Process (Bertsekas, 2011) where state s_t consists of alertness s_t^{alert}, hypertension s_t^{hypert}, intoxication s_t^{tox} time since slept s_t^{tawake}, *total time elapsed* s_t^{ttotal} and *total work done* s_t^{done}.

Over these parameters, we define *productivity* $\eta(s_t^{\text{alert}}, s_t^{\text{tox}})$ and *heart attack risk* $r(s_t^{\text{hypert}})$. The agent receives small positive rewards for productivity and a very large negative reward if a heart attack occurs.

Every half an hour awake, the agent observes s_t and picks an action a_t from discrete action space of *just work*, *drink coffee* (increases s^{alert} and s^{hypert}), *drink beer* (decreases s^{alert}, increases s^{hypert} and s_t^{tox}) and *go to bed* (sleep takes a lot of time, but reduces s^{hypert} and s_t^{tox} and without it alertness starts to fall very fast)

Algorithm	Model	Score	Sleep	Drinks
CEM	0	-524.12		
CEM	3x16	-523.88		
SARSA	0	-130.9	Yes	
SARSA	3x16	-134.95	Yes	
DQN	0	-119.95	Yes	
DQN	3x16	**-84.96**	Yes	
Reference	-	-119.76	Yes	Yes

Table 1: Reinforcement learning compared to reference strategy. All models trained to avoid caffeine and alcohol

3 EXPERIMENTS

We train 2 models (a neural network with 0 hidden layers against one with 3 hidden layers of size 16) with 3 industry-standard algorithms: CEM (Szita and Lörincz, 2006), SARSA (Sutton et al., 1998, Chapter 6) and DQN (Mnih et al., 2013; Mnih et al., 2015) and compare the resulting agents with a reference strategy of sleep every night followed by a cup of coffee.

We train all models with `keras-rl` (Plappert, 2016), limiting all episodes to 1000 steps, and test 20 times. Scores in table 1 are obtained by averaging over total rewards for the 20 test episodes.

We have observed that during first epochs of training the models invariably tend towards high coffee and alcohol consumption (behavior with immediate positive reinforcement but long-term negative side effects), but then converge to more conservative strategies focused on timing sleep correctly and avoiding drinks. This, together with the fact that only one of RL algorithms has outperformed the reference strategy supports our claim that HeartPole is *non-trivial*. At the same time it provides a high level of *transparency*, making it easy to develop and analyze algorithms for reinforcement learning in Healthcare.

Python implementation is available at https://github.com/vadim0x60/heartpole

REFERENCES

Barto, A. G., Sutton, R. S., and Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, (5):834–846.

Bertsekas, D. P. (2011). Dynamic programming and optimal control 3rd edition, volume ii. *Belmont, MA: Athena Scientific*.

Johnson, A. E., Pollard, T. J., Shen, L., Li-Wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Kiani, A., Ding, T., and Henderson, P. rlsepsis234/writeup.pdf at master · akiani/rlsepsis234. https://github.com/akiani/rlsepsis234/blob/master/writeup.pdf. (Accessed on 10/29/2020).

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

Plappert, M. (2016). keras-rl. https://github.com/keras-rl/keras-rl.

Sutton, R. S., Barto, A. G., et al. (1998). *Introduction to reinforcement learning*, volume 135. MIT press Cambridge.

Szita, I. and Lörincz, A. (2006). Learning tetris using the noisy cross-entropy method. *Neural computation*, 18(12):2936–2941.