INTERMEDIATE REWARDS IN REINFORCEMENT LEARNING DRIVEN SEPSIS TREATMENT

CS282r - Semester Project Proposal

LINYING ZHANG[1], DONGHUN LEE[2], SRIVATSAN SRINIVASAN[3]

[1]Harvard University

10 October, 2017

1. Problem Statement

A very common problem in several reinforcement learning applications is the problem of sparse rewards or delayed rewards. In the case of clinical diagnostic treatment, with specific mention to our problem space of sepsis management in ICU, any current literature on solving for the optimal policy obtains rewards only at the end of the treatment, characterized by the mortality of the patient, which makes the rewards matrix both sparse and delayed. We propose to introduce metrics to characetrize intermediate rewards in the treatment trajectory and subsequently verify if the intermediate rewards allow the optimal policy to choose "better"states/actions all along the trajectory and hence obtain better policy values.

2. Challenges

Engineering rewards for different states comes with several layers of challenges. There is no measurable real-time feedback for these abstract rewards unlike the final rewards of death or mortality. Pertinent questions would include if each state deserves a reward or if only salient states do, if the timing of the rewards in the sequencing matter(simplistically could be interpreted as adaptive gamma), if the intermediate rewards model should be purely functions of states or if they are adaptive to patient characteristic features(such as age,gender etc.), if the costs of actions can be treated equally so that the rewards become purely a function of state, if there exist hidden physiological variables beyond the dataset.Besides, understanding intermediate rewards warrant strong clinical interpretation of features to reduce the complexity of the problem space. Also, the feature observations collected during the sepsis case are highly error-prone and poses dangers of noise on any reward construction scheme. With all these challenges and many more involved, we expect to adopt an incremental approach to this problem by building smaller subproblems with very restrictive assumptions and relaxing them iteratively.

3. Outline of Approach

- **PROOF OF CONCEPT and SOFA** : Set up the MDP and propose a few vanilla intermediate reward metrics.This could be asserted on a simple cases like large error-prone

GridWorlds or on the Sepsis dataset itself. This can provide some conviction on the scale and complexity of the problem and provides us some top-level insights into relations between error in intermediate rewards and MDP solutions(by comparing different reward heuristics). We can assume vanilla reward functions such as learning a simple distance measure of the current state from the two absorbing states over the dataset(as a simple supervised learning problem) and assign rewards based on a function of this distance. Our earlier assignments helped verify the strong correlation of SOFA score and mortality. It could serve as a kickstarter to understanding intermediate rewards and we can study the correlations across SOFA and mortality, study sparsity based on states grouping, learn a transformation function based on the correlations and apply them as simple intermediate rewards to solve the MDP. We expect this learning process to be highly noisy as the state action space is large and we have only nominal amounts of data. Also, initially for computational ease, we propose to solve this over a discrete state action space. It could be later scaled to continuous states based on similar literature. **We expect to complete this phase before the next Checkpoint.**

- **FACTOR STUDY AND INFERENTIAL LEARNING**: Once we are able to assert the benefits of intermediate rewards, we conduct a feature study into understanding the most relevant features that have a strong impact on mortality and consequently, our MDP solutions. This could be accomplished by any off-the-shelf feature engineering methods. Once we understand the features, it can provide us a domain under which we can apply several variants of learning methods(both supervised and unsupervised variants) to learn optimal rewards from the features themselves. As an incremental enhancement, we propose leveraging the abstraction power of intermediate feature/state representations learned by deep learning models to effectively infer reward functions. This model could then be trained and tested on slices of dataset. Again, we are going to be fazed with lot of noise in this sample as there is strong disparity in the dataset.
- **CLINICAL INTERPRETATION** : Once we have a stable rewards model, the next step would be to meaningfully interpret these intermediate rewards under the pretext of clinical features and thus create a real-world understanding of what parameters characterize "good"states understood by the optimal solutions of the MDP.This might be impossible too as the abstractions understood by the learning models could be a complex non-linear interplay of clinical features(similar to deep neural networks).
- **FURTHER WORK** : If time permits any further iterations, we can study this as an inverse RL problem to understand the interpretation of clinician policies. Another interesting addition to this space would be to study the temporal component of these intermediate rewards,as to how they vary for the same states based on the time bloc they occur in patient history.

4. Evaluation

At different stages of the problem, we are required to evaluate the success of the proposed reward scheme. We propose a simple comparison of solving the MDP without any intermediate rewards and compare the value of the optimal policy learned with and without the presence of intermediate rewards. It would also be pertinent to study cases(in terms of both patient features and trajectories) under which the intermediate rewards add value to the MDP solution.

5. Initial References

While drafting the initial proposal, we surveyed the following references which span across reinforcement learning theory, robotics and clinical treatment domains.

- https://arxiv.org/pdf/1612.06699.pdf
- http://ieeexplore.ieee.org/document/6420413/
- http://ieeexplore.ieee.org/document/7578592/
- Raghu et al. Continuous State-Space Models for Optimal Sepsis Treatment - a Deep Reinforcement Learning Approach.
- Springer Handbook of Robotics