R6/Ishank Juneja/16D070012

February 8, 2020

While treating chronic disorders, clinicians need to make decisions about the sequence of treatments for their patients. Some chronic diseases affect a significant number of people. In particular, epilepsy is a chronic neurological disorder that affects nearly 1% of the population. The disease is characterized by recurrent seizures and sensory disturbances. Treatment for epilepsy involves stimulation of the brain by the means of implanted electrodes. The specific treatment administered to a patient can be boiled down to a stimulation pattern programmed into the implant.

These administered patterns can be modelled as policies for an underlying sequential decision making task. The problem tackled by the paper is to find a near optimal policy for the stimulation administration task. The authors demonstrate how reinforcement learning coupled with a suitable function approximator can be used to achieve this goal.

The state for this problem consists of continuous valued data coming from an EEG (electro-encephalogram) reading. Since the sequence space is infinite dimensional, features are extracted from the EEG data. As described in the paper, extracted features consists of *frequency domain* information about the sequence of EEG values. This choice of features seems intuitive since the decisions made by the algorithm involve providing periodic stimulations at different frequencies. In particular the four possible actions are stimulation at $0.2Hz$, $0.5Hz$, $1.0Hz$ or no stimulation at all. To produce desirable results, the optimal policy will need to walk a fine line between too much and excessive stimulation. The reward scheme used for the problem is reflective of this fact.

Since the treatment problem involves stimulation of a sensitive organ learning online is not practical. The paper proposes to apply *Batch-RL* to train the agent on pre-recorded trajectories. The action value function for the underlying MDP - $Q(s, a)$ - is estimated using the Fitted Q-Iteration algorithm. The choice of Extremely Randomized Trees (Extra-Trees) is made as the function approximator for the action values.

The authors describe four scoring metrics for empirical evaluation. These metrics include - proportion of states spent in seizure, No. of stimulations used and expected return. Empirical evaluation of the policies learnt from the action value functions show a tremendous advantage of using RL based strategies when compared to the traditional strategy of choosing single action over test data.

Some comments/questions I have are -

- The problem tackled by the paper can be summarised as choosing actions based on the history of a temporal sequence. Keeping this in mind, how does the paper's performance compare to approaches like *Hidden Markov Models* etc. that cater to sequence prediction?
- I found it curious that the authors compare *Extra-Trees* to feed-forward neural networks which, as per my understanding, are not considered to be rich enough for learning from temporal sequences.