OPTIMAL BIOPSY DECISION MAKING IN BREAST CANCER USING REINFORCEMENT LEARNING

By

Ruchi Parmar
501034872

Methodology & Experiments

Master of Science
in the Program of
Data Science and Analytics

Toronto, Ontario, Canada, 2024

Table of Contents

1 Methodology and Experiments **1**

A Aim of Study 1

B Response(Dependent) and Independent Variable(s) 1

C Factors and Levels 1

D Experimental Design 1

E Experiment Performance and Revisions 4

F Measuring Performance 4

G Algorithm Comparison and Selection 4

1. Methodology and Experiments

A. Aim of Study

The aim of this study is to develop and evaluate reinforcement learning (RL) algorithms to optimize breast cancer screening and diagnostic decision-making. Specifically, the study focuses on determining the optimal actions(Annual Mammogram or Biopsy) at different risk levels and ages to maximize long-term health outcomes and minimize unnecessary procedures. different RL approaches like traditional MDP based Backward Induction and advanced method, Q-learning is implemented for the study.

B. Response(Dependent) and Independent Variable(s)

- **Response(Dependent) Variable**:
  - Long-term health outcome, measured by the reward function which incorporates the health states(Healthy, Early, Advanced, and Terminal) and the utility of diagnostic actions
- **Independent Variables**:
  - Age of the women(ranging from 40 to 99 years)
  - Risk score of the women(a metric representing the probability of having breast cancer)

C. Factors and Levels

In this experiment, the factors are the different RL methods being implemented and tested. Levels for the experiments are the optimization of Biopsy decision making while reduce unnecessary procedures, and mitigate the economic impact of false positives.

D. Experimental Design

The study employs two RL methods to develop and compare policies for optimal decision-making - Backward Induction and Q-Learning. The problem is framed using concepts of - Markov Decision Process(MDP), which are -

- **States(S)**: Represent the health status of a woman, captured by the risk score and categorized into different health states(Healthy, Early-stage Cancer, Advanced-stage Cancer, Terminal)
- **Actions(A)**: Possible actions, namely Annual Mammogram(AM) and Biopsy(B)
- **Transition Probabilities(P)**: Probabilities of moving from one state to another given a specific action
- **Rewards(R)**: Immediate rewards received after transitioning from one state to another, reflecting the health outcome and cost of actions

RL approaches implemented -

1. **Backward Induction**: Backward induction is a dynamic programming method used to solve MDPs by iteratively computing the optimal policy starting from the terminal state and working backward to the initial state.

   Algorithm Steps:

   (a) **Initialization**:
   - Define the set of states(S), actions(A), and time steps(T)
   - Initialize the utilities(U) for each state at the final time step(age 100) based on terminal rewards

   (b) **Backward Iteration**:
   - For each time step t from 99 to 40:
     - For each state s in S:
       * Compute the expected utility for each action a in A by considering the rewards and the expected future utilities
       * Select the action that maximizes the expected utility
       * Update the utility of state s and the policy for time step t

   (c) **Policy Extraction**:
   - The resulting policy indicates the optimal action to take at each state and time step

   States:

- Risk scores ranging from 0 to 100
- Terminal state 100, which is an absorbing state indicating cancer detection

Rewards:

- Defined for each combination of state, action, and time step
- Example: Reward for 'AM' and 'B' actions in Healthy/ Early-stage cancer/ Advanced stage cancer state at given time stamp(age) t

2. **Q-Learning**: Q-learning is an off-policy temporal difference learning method used to find the optimal action-selection policy for an MDP by learning the action-value function Q(s,a) through interactions with the environment.

   Algorithm Steps:

   (a) **Initialization**:
   - Initialize Q-values for all state-action pairs to zero
   - Set the learning rate(alpha) and discount factor(gamma)

   (b) **Epsilon-Greedy Policy**:
   - Implement an epsilon-greedy policy to balance exploration and exploitation. With probability epsilon, select a random action; otherwise, select the action with the highest Q-value.

   (c) **Action Execution**:
   - For each episode:
     - Start from the initial state
     - At each time step, select an action based on the epsilon-greedy policy
     - Execute the action, observe the reward and the next state
     - Update the Q-value for the state-action pair using the Bellman equation:
     - Transition to the next state and repeat until the terminal state is reached or the episode ends

   (d) **Policy Extraction**:
   - The optimal policy is derived by selecting the action with the highest Q-value for each state.

   States: Same as backward induction, ranging from risk scores 0 to 100

   Rewards: Same as backward induction, defined for each state-action pair

Exploration vs. Exploitation: The epsilon-greedy strategy ensures that the model explores new actions with probability epsilon and exploits the best-known actions with probability(1 - epsilon).

E. Experiment Performance and Revisions

The experiments are conducted in multiple iterations to refine the RL algorithms and improve the policies:

1. **Initial Setup**: Implement backward induction and Q-learning algorithms using the datasets. The initial methods are run with basic hyper-parameters and without any optimizations.
2. **Revisions**: Based on the initial results, several revisions will be made:
    - Adjust the learning rate(alpha) and discount factor(gamma) for Q-learning
    - Refine the reward function to better capture the clinical significance of different health states
    - Add exploratory strategies(epsilon-greedy) to balance exploration and exploitation in Q-learning

F. Measuring Performance

Policy Evaluation: The performance of the RL methods are measured by evaluating the learned policies.

G. Algorithm Comparison and Selection

The following criteria will be used to compare the backward induction and Q-learning algorithms:

- **Convergence**: Speed and stability of convergence to an optimal policy
- **Policy Quality**: Effectiveness of the policy in maximizing long-term rewards and health outcomes
- **Computational Efficiency**: Time and resources required for algorithms to get the policy