**A Reinforcement Learning Method Based on Information Evaluation**

Shuai Wan

**Abstract**

The success of current machine learning techniques is mainly attributed to a large amount of independently and identically distributed (i.i.d.) training data. However, due to the inability to effectively understand causal relationships, they exhibit poor environmental transferability. In the process of understanding and transforming the world, humans are good at summarizing and abstracting knowledge from environment. This abstract knowledge can easily adapt to environmental changes, greatly improving the scope of knowledge application. In order to imitate this abstract ability of humans, this paper proposes a model-based reinforcement learning method that builds a information-relation model of the environment, continuously assumes the correlation between various information in the environment, and observes the size of the gain brought by various hypotheses in interaction, so as to continuously optimize the model and mine the information-relation model of the environment. I tested it in a self-made mine swap game, and the experimental results show that the modeling results of the model can guide the agent to make correct decisions in different changes of the game, and it is easy to transfer its modeling results to similar environments.

**1. Introduction**

Learning causal relationships has long been a milestone challenge in artificial intelligence. Humans can intuitively infer causality between elements without explicit guidance. Most current machine learning successes rely heavily on pattern recognition from large, well-collected independent and identically distributed (i.i.d.) datasets. However, in real-world scenarios, distributions often shift due to unaccounted or uncontrollable factors in training data. For machine learning models to function beyond i.i.d. domains, they must go beyond learning statistical correlations between variables and instead uncover underlying causal models [1].

This paper argues that integrating causal models into supervised learning has significant limitations. First, the added causal models cannot evaluate or optimize causality itself. Second, supervised learning relies on domain-specific knowledge, which is inherently narrow—capturing only a few causal relationships within that domain, making it difficult to compare or derive new, higher-order relationships from a broader set.

Reinforcement learning (RL) can effectively address these shortcomings. First, RL generates experimental results through continuous interaction with the environment, theoretically providing an infinite sample pool. Second, RL can incorporate specific search strategies to prioritize directions deemed important by the policy, enhancing sample efficiency. Third, in the RL algorithm proposed here, the search strategy evolves based on a continuously expanding and optimized information-relation model, allowing sample efficiency to improve progressively throughout the learning process.

**2. Background**

The Markov Decision Process (MDP) is the foundation of modern reinforcement learning. RL algorithms model the exploration of an environment as transitions between state-action pairs, aiming to converge on an optimal action or state-value function. However, MDP-based RL overly emphasizes objective state transitions, neglecting the subjectivity of decision-making. This limits RL to merely learning state-action mappings without exploring why a specific action is taken in a given state.

This paper proposes a method that leverages information observed by an agent to hypothesize relationships between data points and validates these hypotheses through environmental interaction. This approach extends RL beyond optimizing a value function—its goal becomes uncovering as many intrinsic relationships within the environment as possible.

1. **Algorithm**: A Reinforcement Learning Method Based on Information Evaluation

while True:

1. Generate new hypotheses based on existing information.

2. Play multiple rounds, pruning and simulating based on these hypotheses.

3. Use Monte Carlo Tree Search (MCTS) to calculate hypothesis hit rates and evaluate the effectiveness of decision-making behaviors.

4. Discard hypotheses with hit rates close to random; retain high and low performers for further evaluation.

5. Invert low-performing hypotheses, preserve high-performing ones, and analyze surrounding information (downward, peer-level, and upward observations).

6. Generate additional hypotheses (based on existing information), explore their surroundings, summarize relationships, and propose new hypotheses (downward, peer-level, upward).

If the expected reward approaches the optimal model:

Break;

Example (Eating an Apple for Reward):

Downward Observation: When evaluating “apples are edible,” we examine sub-hypotheses like “green apples are edible” and “red apples are edible,” confirming their validity.

Peer-Level Observation: Upon confirming apples are edible, we note apples belong to the “fruit” set (alongside oranges and bananas) and hypothesize whether oranges and bananas are edible.

Upward Observation: With many positive results, we hypothesize upward: “Are all fruits edible?” based on prior experience.

To demonstrate this information-relation model’s transferability and adaptability, I implemented it in a minesweeper game simulation, varying board sizes and mine placement rules to test the model’s flexibility.

**4. Minesweeper Game**

This is a custom minesweeper game I designed. The rules are as follows: at the start, the top-left cell displays a number indicating the count of mines in its four adjacent cells (up, down, left, right). Each column contains exactly one mine. The game ends when all non-mine cells are cleared. Stepping on a mine yields a -1 reward. The board size is customizable.

Clearly, the more information a player has, the easier it is to win. Information falls into three categories:

1. Visible board data.
2. Inferences from numbers and their combinations.
3. Hidden rules (e.g., one mine per column) discovered through multi-round simulation.

Using Monte Carlo random simulations, we can estimate the expected reward given the current information level and the marginal reward each piece of information contributes. Information proven consistently useful is stored in the model, becoming part of the agent’s learned knowledge.

**5. Experimental Results**

On a standard CPU (AMD Ryzen 7 3700X 8-Core Processor), I ran the simulation for 3 minutes, completing 1,000,000 games. By hypothesizing one-hop relationships (e.g., “if the number is 2, the bottom-left cell must contain a mine”), the model extracted 71 relationships (stored in edr.json). This reduced the game failure probability from approximately 0.86172 to 0.773, providing preliminary evidence of the algorithm’s convergence.

([The code is available on GitHub](https://github.com/qqqqqwwwweeerr1gmail/gitShare/blob/main/code_my/mine_swap/20220701/code_A%20Reinforcement%20Learning%20Method%20Based%20on%20Information%20Evaluation.py))

**6. Conclusion**

MDP-based reinforcement learning focuses on learning state and action value functions, requiring retraining with even minor environmental changes. In contrast, reinforcement learning based on information evaluation learns intrinsic relationships between information, offering strong adaptability to environmental shifts. This broadens the applicability of knowledge beyond specific contexts. The extensive connectivity of knowledge hints at a pathway toward Artificial General Intelligence (AGI).

**References**

[1] Dickson, Ben. "[Why Machine Learning Struggles with Causality](https://bdtechtalks.com/2021/03/15/machine-learning-causality/)."