論 文 の 内 容 の 要 旨

論文題目　Application of Simple Regret Bandit Algorithms on Monte-Carlo Tree Search
（単純リグレットバンディットアルゴリズムを利用したモンテカルロ木探索）

氏　　名　劉　雲青

Monte-Carlo Tree Search (MCTS) has made a significant impact on various fields in AI, especially on the field of computer Go. The key factor to the success of MCTS lies in its combination with bandit algorithms, which solves the multi-armed bandit problem (MAB). The MAB problem is a problem where the agent needs to decide whether it should act optimally based on current available information, or gather more information at the risk of suffering losses incurred by performing suboptimal actions. One of the most widely used MCTS variants is the UCT algorithm, which simply applies the UCB algorithm to every node in the tree.

The pure exploration MAB problem seeks to identify the optimal best arm, rather than gathering as much reward as possible. The pure exploration MAB problem can also be formally stated as the minimization of simple regret, which is defined as the difference between the expected reward of the optimal bandit and the bandit that has been identified as the optimal bandit. Since the main objective of game tree search is to identify the best action to take, it has been considered that bandit algorithms that solve the pure exploration MAB problem would be a better match for application in MCTS. However, the application of simple regret bandit algorithms to MCTS is far from trivial.

The simple regret bandit algorithm has the tendency to spend more time on sampling suboptimal arms, which may be a problem in the context of game tree search. In this research, we will propose combined confidence bounds MCTS (CCB-MCTS) algorithm, which utilize the characteristics of the confidence bounds of the improved UCB and UCBsqrt algorithms to regulate exploration for simple regret minimization in MCTS.

Another possible approach is based on the observation that max nodes and min

nodes in game trees have different concerns regarding their value estimation, and different bandit algorithms should be applied accordingly. We develop the Asymmetric-MCTS algorithm, which is an MCTS variant that applies a simple regret algorithm on max nodes, and the UCB algorithm on min nodes.

Both the performance of the CCB-MCTS and Asymmetric-MCTS algorithm has shown good performance on the games of 9x9 Go and 9x9 NoGo. The empirical performance of the Asymmetric-MCTS algorithm also revealed the effectiveness of the applying simple regret bandit algorithm seems to be related to the structure and distribution of the values at the leaf nodes of the game tree.