Improvements of Penalty Avoiding Rational Policy Making Algorithm used proof number and disproof number

Takashi Nobori (310083)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 10, 2005

Keywords: reinforcement learning, Penalty Avoiding Rational Policy Making Algorithm, number of proof(disproof), neural network, othello.

Reinforcement learning is a kind of machine learning system. It aims to adapt an agent to a given environment with a clue to rewards. In general, the purpose of reinforcement learning system is to acquire an optimum policy that can maximize expected reward per an action. Thus, the agent can automatically acquires the optimal policy without an explicit supervisor, if the reward function is defined.

When we aim to construct a policy (a mapping from faces of a board to moves) in a board game, sufficient knowledge about the game are usually required and must be introduced to an evaluation function for the faces of the board. On the other hand, when applying reinforcement learning to it, we need to define only the reward function on the state of the win, and then an agent is expected to acquire a policy to reach a win of the game through the learning. There are several applications of reinforcement learning to the board games. For example, TD(λ) was used to learn the neural network expressing the faces of the board of the Othello. In another example, TD(λ) is used to adjust the weights of the evaluation function, that is the evaluation function used in the "Knight Cap", chess player program. These approaches aimed to obtain optimal policies, the policies that lead to a great victory.

However, in 2-players game such as the Othello game, it is more important to acquire a penalty avoiding policy than a optimal policy. Miyazaki et al. proposed "Reinforcement learning for penalty avoiding rational policy making (PARP)" method for the purpose of applying reinforcement learning to such problems and applied it to the Othello game. This algorithm try to form the policy which avoid the states that lead to the lose of the game by memorizing such states experienced before. In this algorithm, a penalty state is defined as the states from which a player always lose. The penalty states are transmitted from the end states to the start states of the game if all moves that can be chosen in a state result in the movements to the penalty states. However, when applying this algorithm to the large-scale problem such as a Othello game, we apparently confront the curse of dimensionality. Especially, a huge number of trials must be requited to transmit penalty states in the middle stage of the game, because the agent can not receive the reward until it reaches the end states of the game. To overcome these problems, Miyazaki et al. proposed some ideas. They restricted the expansion of the nodes from the current node in the search tree needed to judge whether the current node is penalty one or not. They also introduced the additional semi-penalty, which is defined according to knowledges about the Othello games, and accelerated the transition of the penalty states in the middle stage of the games. The semi-penalty is an indicator that indicates the certainty that the node is really a penalty node. However, the performance of the PARP method is strongly dependent on the degree of accuracy of the semi-penalty.

In this research, we introduce the proof number and the disproof number into the original PARP algorithms to transmit the penalty state more efficiently. The proof number is defined by the minimal number of the leaf nodes ,that have already expanded from the current node and must be proved to be win nodes, needed to confirm that the current node is a win node. And the disproof number is also defined by the lose. These number can be used as indicators of estimating the difficulties of identifying win or lose on a node. So, we use these indices to give priorities of the expansion of the nodes needed to transmit penalty states. By using the depth-first search according to the disproof number, it is expected that the the penalty state can be transmitted more efficiently. Consequently, we can reduce the

dependency on the heuristic semi-penalty.

In the experiments, we apply the proposed method to the Othello game. As the opponents, we designed two type of players which select moves according to the α - β method using the well- known heuristic evaluation functions. The level of the opponents can be changed by altering the depth of the α - β search. As a result of the comparisons, we show that the proposed method can find the policy which defeat the several levels of the opponents with the smaller number of trials than those needed for the original PARP. Especially, the proposed method can find such policy under the situation that the dependency on the heuristic is so small that the original PARP can not defeat the opponent.