Title	強化学習における危険回避行動獲得のための負の報酬伝搬法
Author(s)	寺田，賢二
Citation	
Issue Date	1998-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1129
Rights	
Description	Supervisor:國藤 進，情報科学研究科，修士

Exploitation-Oriented Reinforcement Learning with the Aim to Avoid a Penalty

Terada Kenji

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

February 13, 1998

Keywords: reinforcement learning, classifier system, bucket brigade algorithm, penalty avoidance.

Abstract

Assuming that a learning agent such as the autonomous robot is worked at a complicated environment or a dynamic environment, it is hard for a robot designer to design a robot which adapt to the environment. Therefore, the agent needs an adaptive ability. In the case of autonomous robot learning, since the robot is a high price, the penalty avoidance is as important as the efficient task. Therefore, the purpose of this paper is to acquire rule of the penalty avoidance by reinforcement learning.

The leaning by reinforcement learning reinforcement the rule given positive reward. Therefore, many studies of reinforcement learning don't regard the penalty avoidance. The learning of penalty avoidance reinforces the only relatively increase of strength to reduce the strength of other penalty rules. This means that rule of penalty avoidance is judgment an invalid rule. Therefore, the course of this study is to make rules of penalty avoidance of valid rules.

The proposing method of this paper uses classifier system. Classifier system is a kind of exploitation-oriented reinforcement learning. Bucket brigade algorithm is the mechanisms for apportionment of strength of classifier system. The proposing method improves bucket brigade algorithm. The alteration is to add the process for directly rewarding a rule of avoidance penalty.

A set other reward besides a set reward for an environment obstructs an acquisition of reward for an environment because learning agent is satisfied to gain other reward. Therefore, the proposing method solve this problem by an equivalent negative propagation to reward of the penalty avoidance. But it is hard for negative propagation to select on the rule which apportion negative strength. To solve this problem, the proposing method

doesn't apportion a strength of one rule, but strengths of all activated rule. It is not necessary to select by this process.

The rule abolition in the rule generation process doesn't select the rule having the least strength, but the rule having the least absolute value of the strength , because negative propagation uses negative strength. In the case where the rule generating is only the genetic algorithm, it is impossible to generate rule while reinforcement learning. The question now arises: In a state without a active rule, the learning agent is impossible to decide on a policy. Assuming that learning agent decides random behavior in this state, not only the learning agent spoils the learning opportunity; In the case where this state includes behavior of penalty, the learning agent is usual in danger until the genetic algorithm practice. Therefore, the learning agent in this paper also generates a rule when all rule doesn't active.

We verify the ability of the proposing method by experimenting on three simulation with six mechanisms for apportionment of strength.

The six mechanisms for apportionment of strength is as follows:

1. General bucket brigade algorithm

2. The method is added negative propagation to method 1.

3. The method is added reward for avoidance to method 2.

4. The method is added the propagate strengths of all activated rule to method 1.

5. The method is added the propagate strengths of all activated rule to method 2.

6. The proposing method (The method is added the propagate strengths of all activated rule to method 3)

The three simulation and those characteristics are as follows:

Garbage collection problem : A set reward for an environment exists at random. A sequence of behavior for a set reward for an environment is not the loop.

Maze learning problem : A set reward for an environment exists a dangerous zone. A sequence of behavior for a set reward for an environment is the loop.

Missionaries-and-cannibals problem : A set reward for an environment exists beyond a dangerous zone. A sequence of behavior for a set reward for an environment is the loop

The following results were obtained:

- The proposing method is punished smaller than general bucket brigade algorithm in all problems. We may, therefore, reasonably conclude that the proposing method has the ability to reduce penalty.

- The proposing method obtains the number of a set reward for an environment smaller than general bucket brigade in garbage collection problem but as much as general bucket brigade in maze learning problem and missionaries-and-cannibals problem. and the proposing method also obtains as much as general bucket brigade in garbage collection problem by adjusting reward for avoidance.

- The method is added the propagate strength of all activated rule obtains a set reward many more than others. These results lead to the conclusion that there is a possibility that this method improve the search ability.