203 A Study of Opportunity Evaluation of Reinforcement Learning Method

Hokkaido University ○Yu Wenwei Kakazu Yukinori

Abstract

There existed the research effort using the experience to guide the exploration of the state-space in the area of Reinforcement Learning. In this research, the metric and methods used to evaluated the opportunity of the previous experience to reduce the searching and learning time are studied, and incorporated with the Q-Learning and DP like method. By solving the large size 3-DOF problem, we show the efficiency of each evaluation and suggest the promising one.

1. Introduction

Experience reuse via Priority queue[1][3] has been favored by some researches to cope with the slowly convergence of some well known RL methods, such as Q-Learning and DP like method. In this research, effects have been done to the opportunity evaluation, i.e., how to decide the current step is interesting, and what can be done, therefore to improved the performance of RL methods.

2. Priority queue and Opportunity Evaluation

The experience reuse via priority queue can be abstracted to the model is displayed in Figure 1. It can be noticed that, the step2, step3 play an important role in the whole process. They will be emphasized in the next discussion of this section.

```
Queue Building and Maintenance for each A (action) in S (state):
Until Queue_Length == 0 or Backup_Number > Predefined_Number
        Step1: Expected_Return_Calculation
        Step2: Function_Opportunity_Metric_Calculation( Variables: Qold, Qnew, and/or  Others )
        Step3: IfInteresting = Opportunity_Evaluation( Variables: Opportunity_Metric, Threshold and/or Others )
        Step4: if IfInteresting == INTERESTING
                        if the pair (Current_State, Opportunity_Metric, NextState ) already in the queue
                                promote the pair
                        else add the pair to the queue
```

Figure 1 Experience reuse through priority queue

2.1 Using expected λ discounted return as the metric

Using the expected λ discounted return as the the estimator of Q value can utilize more information about thesearch process

2.2 Using the old Q-Value as the threshold

We regard that using certain Δto decide if interesting, will lead to the over estimating or miss estimating the current opportunity, whereas using old Q-value can keep consistence with the essence of the temporal difference theory and decide the situation well.

```
Label Process:
        if NextState == StartState
                return LABEL_WORSE
        else if Label(NextState) ==UNLABELLED
                Label(NextState) = Label(CurrentState) + 1
                return LABEL_UNIMPROVED
        else if Label(NextState) > Label(CurrentState) + 1
                Label(NextState) = Label(CurrentState) + 1
                return LABEL_BETTER
        else if Label(NextState) == Label(CurrentState) + 1
                return LABEL_UNIMPROVED
        else
                return LABEL_WORSE
Opportunity Evaluation:
        if RETURN == LABEL_BETTER
                do Lable_Refresh
                if Q(CurrentState) != Init_Value
                        do Q_Value_Refresh
```

Figure 2 Label Process and Opportinity Evaluation

2.3 Using step labelling method as the evaluation scheme.

Further more, we suggest the more direct metric and evaluation scheme: labelling each state with the step number from the start step, if the current action will lead to reduce the step number, then interesting(Fig 2).

3. Problem Setting and Experiment Results

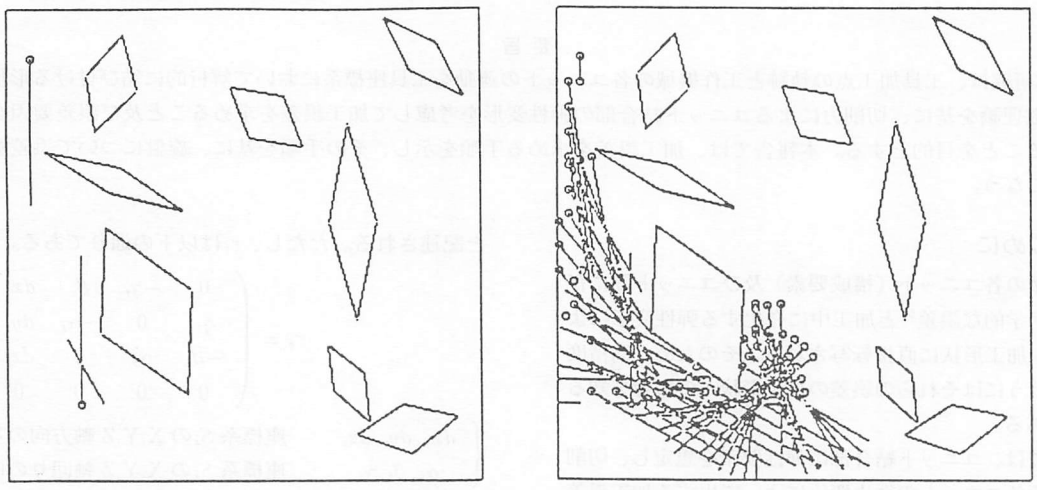

Figure 3 3-DOF problem and one solution

3-DOF Problem and one of the solution is displayed in the Figure 3. We set the area as 20x20, and rotate interval is 10 degree. α=0.3, β=0.3, λ=0.2, n=ProbActionNumber.

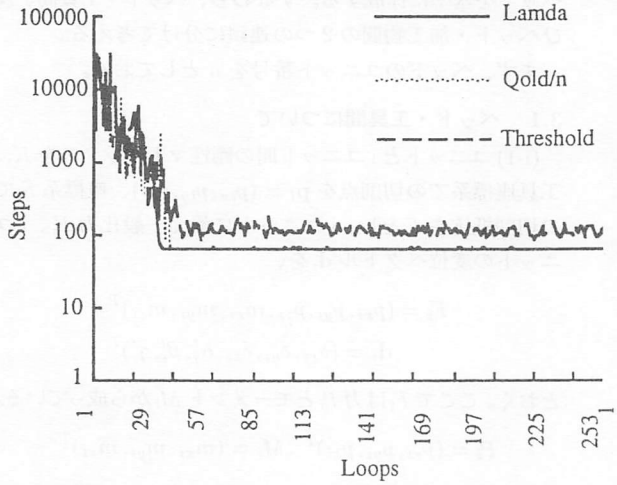

Figure 4 comparing between evaluation based on Q(λ), Qold and threshold

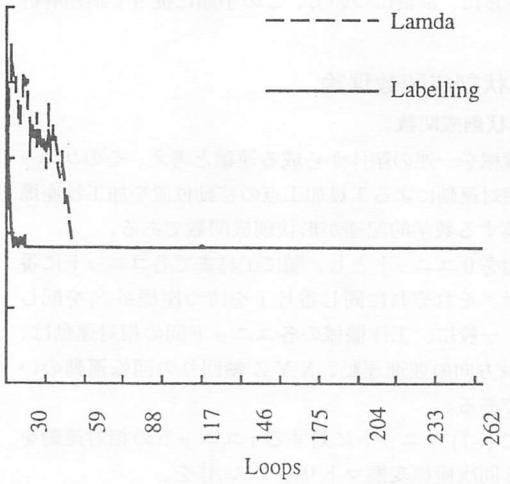

Figure 5 comparing evaluation based on Labelling with Q(λ)

As analyzed in section 2, the Qold/n based method can perform over the certain threshold based method, while the Lamda based evaluation pay off the quick convergence with expensive calculation. The Labelling based evaluation can begin the learning from first experience, can perform well with not too high cost.

4. Conclusion

Aiming at the efficiency of the RL methods, we studied some elements of opportunity evaluation for experience reuse, and suggested the step labelling based evaluation which actually add more heuristics to the learning method. The performance improvement is shown through experiment.

Reference:

1) Andrew W. Moore, Christopher g. Atkeson, Memory-Based Reinforcement Learning: converging with less data and less real time, in Robot Learning, edited by Jonalthan H. Connell and Sridhar Mahadevan, 1993, Kluwer Academic Publishers.

2) Sutton, R. S., Learning to predict by the method of temporal differences, Machine Learning 1988, pp9-44.

3) Jing Peng and Williams R. J., Efficient Learning and Planning Within the Dyna Framework, From Animals to Animats 2.