Reinforcement learning approach to multi-stage decision making problems with changes in action sets

Takuya Etoh[1], Hirotaka Takano[1], and Junichi Murata[1]

[1] Kyushu University, 744 Motooka, Nishi-ku, Fukuoka, Japan
(Tel: +81-92-802-3675, Fax: +81-92-802-3692)
(etoh@cig.ees.kyushu-u.ac.jp)

Abstract: Multi-stage decision making (MSDM) problems often include changes in practical situations. For example travelling time of a path changes in the path selection problems in road networks. The changes cause risks in adopting solutions to MSDM problems. Therefore, we propose a method for solving MSDM problems considering risks. Reinforcement learning (RL) is adopted as a method for solving those problems, and stochastic changes of action sets are treated. It is necessary to evaluate risks besed on subjective views of decision makers (DMs) because the risk evaluation is by nature subjective and depends on DMs. Therefore, we develop an RL approach to MSDM problems with stochastic changes in sets of alternative actions, which uses new method for evaluating risks of the changes. The effectiveness of the method is illustrated with a road network path selection problem.

Keywords: Multi-stage desicion making, Risk assessment, Subjective occurrence probability, Reinforcement learning

1 INTRODUCTION

In practical situations, there are various Multi-stage decision making (MSDM) problems with changes. For example, in the path selection problem of road network, travelling time of paths changes. There are several optimization methods to solve MSDM problems without changes. In this paper, we adopt Reinforcement Learning (RL), which needs no models of problems for solving MSDM problems [1], and modify RL to MSDM problems with changes.

When MSDM problems include changes, we have to consider risks caused by the changes. In general, risks are evaluated by the occurrence probabilities of changes and costs incurred by the changes. Here, a cost represents the degree of undesirable effect; it does not necessarily indicate an amount of money. In RL, effects of stochastic events are originally evaluated by the expectation, but risks evaluated by the expectation may not fit subjective views of decision makers (DMs). Therefore, a method is necessary for evaluating risk that can incorporate DM's subjective views.

An RL method was reported for solving MSDM problems with changes of state transition probabilities [2] and RL methods for solving MSDM problems with changes of rewards [3, 4]. The former considers states called error states, which are undesirable or dangerous states, and defines risks as probabilities that the agent enters error states. The latter defines risks as variances of rewards or worst-cases of reward changes. Those adopt a weighted sum of expectations of rewards and costs incurred by changes as value functions. In those methods, adjustments of weights reflect DM's subjective views. The determination of the weight values cannot be done intuitively and therefore is not easy. In contrast to those two types of changes, there are also changes of action sets in practical problems. Therefore this type of changes is dealt with in this paper.

The above-mentioned methods need estimates of costs incurred by changes. However, for changes of action sets, estimates of costs incurred by changes cannot be obtained until the learning converges sufficiently, because costs are only evaluated by the difference between the value of the optimal action and the value of the suboptimal action. Therefore, we focus on probabilities at which action sets change, and propose a new method for evaluating risks with DM's subjective views in the form that can fit the RL framework. We assume that probability distributions are known as also assumed in the above-mentioned articles [2, 3, 4].

2 REINFORCEMENT LEARNING

RL is a method for getting optimal action selection policy automatically based on trial-and-error in order to solve complex or unknown problems. Signals used in RL are denoted as follows:

- discrete-time : $t\ (= 0, 1, 2, \ldots)$,
- state of environment at time t : $s_t \in S$,
- action taken by agent at time t : $a_t \in A_{s_t}$,
- reward obtained from environment at time t : r_t,
- policy of agent : $\pi(s, a)$.

In RL, learning is performed by the interaction between the environment and the agent as shown in Fig.1.

First, after the agent takes action a_t at time t, the state s_t transitions to s_{t+1} at next time $t+1$ by the action a_t. As a result, the agent gets the reward r_{t+1}. Second, the agent updates its own policy π, that is strategy of action selection, based on the obtained reward. The agent takes action based on the policy.

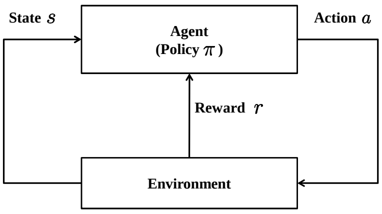

Fig. 1. The interaction between the environment and the agent.

In Fig.1, one cycle is called one step, and sequences of actions and state transitions from starting states to termination states are called an episode.

Because the reward is the only criterion in RL, the agent evaluates merits of actions and states based on the rewards, and those merits are called the values. The action value function is defined as

$$Q^{\pi}(s,a) = E\left[\sum_{k=0}^{\infty} \gamma^k r_{t+1+k} \mid s_t = s, a_t = a, \pi\right]. \quad (1)$$

The equation (1) represents the value of the action $a_t = a$ in the state $s_t = s$ at the time t under the condition that the action is selected by the policy π at time $t+1$ and thereafter. The value is expressed by the expectation of weighted sum of rewards obtained from time $t+1$ to the future.

3 PROPOSED METHOD

3.1 General Problem Setting

We assume Markov decision processes (MDPs). The MDPs include changes of action sets: some of alternative actions become unavailable at certain probabilities in some states. Moreover, there are dependences between occurrences of a number of action changes, and those dependences are described by the conditional probabilities. In this paper, we assume that all of those probabilities are known. Since probabilities that new unknown alternative actions appear are explicitly zero, they do not appear.

3.2 Subjective Occurrence Probability

Risks are evaluated, in general, by occurrence probabilities of changes and costs incurred by changes. When the optimal action becomes unavailable by the change, we have to use a suboptimal action. Therefore, costs are described by difference between the value of the optimal action and the value of the suboptimal action. Occurrence probabilities of changes are assumed to be known, but the costs cannot be obtained until the learning converges sufficiently. Therefore, we consider risk evaluations based on only occurrence probabilities.

Now, consider people who decide to take an umbrella or not after knowing the rainfall forecast. Some people take an umbrella if the rainfall probability is 50 %, but others do not take an umbrella even if the rainfall probability is 50 %. It means that people who take an umbrella consider that it may rain if the rainfall probability is equal to or higher than 50 %, and people who do not take an umbrella consider that it will not rain even if rainfall probability is 50 %. In this way, interpretations of occurrence probabilities vary from person to person.

In this paper, we propose a function $F(p)$ below that can approximate the relationship between objective occurrence probability p and subjective occurrence probability P, and use P as subjective evaluations of the risks,

$$P = F(p) = \frac{1+\exp(\frac{\tau-1}{\sigma})}{\exp(\frac{\tau}{\sigma})-\exp(\frac{\tau-1}{\sigma})}\frac{\exp(\frac{\tau}{\sigma})-\exp(\frac{\tau-p}{\sigma})}{1+\exp(\frac{\tau-p}{\sigma})}, \quad (2)$$

where $\tau(0 \leq \tau \leq 1)$ and $\sigma(\sigma \leq 1)$ are parameters reflecting DM's subjective views. Because $F(\tau) = 0.5$ holds, DMs consider that changes will likely to occur if $p \geq \tau$ and consider that chages will not likely to occur if $p < \tau$, that is, the parameter τ means the threshold value. The parameter σ controls the gradient around the threshold value. The curves of equation(2) with various parameter are shown in Fig.2.

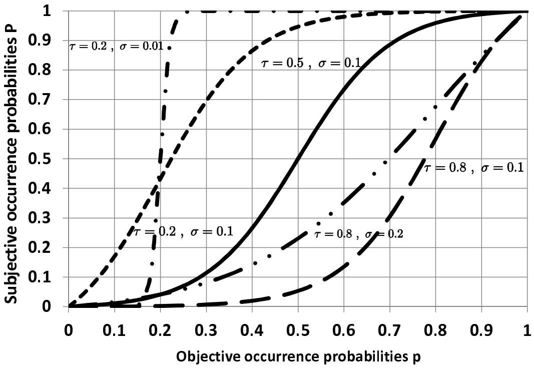

Fig. 2. Subjective occurrence probabilities for objective occurrence probabilities (Function $F(p)$).

3.3 The Compatibility Between the Learning Algorithm and the Risk Evaluation

The value of an action that may become unavailable should be reduced based on DM's subjective interpretations of occurrence probabilities. We multiply the value by a factor $1 - P_t$, where P_t is the subjective probability that the action at time t, a_t becomes unavailable, that is, if the DM interprets that the action will not be available at a high subjective probability ($P_t \simeq 1$), its value is reduced almost to zero. Therefore, the updating equation of Q is defined as

$$\begin{aligned} Q(s_t, a_t) \leftarrow & (1-\alpha)Q(s_t, a_t) \\ & +\alpha(1-P_t)\{r_{t+1} + \gamma \max_a Q(s_{t+1}, a)\}. \end{aligned} \tag{3}$$

This update formula converges to

$$(1-P_t)\sum_{t=0}^{\infty}(\prod_{j=1}^{k}\gamma(1-P_{t+j}))r_{t+1+k}$$

which is a sum of rewards discounted by not only γ but also $1 - P_{t+j}$.

The objective occurrence probabilities p are assumed to be known. But, even if the agent estimates occurrence probabilities, this method is efficient because the convergence of objective occurrence probability estimates is faster than the convergence of value functions.

3.4 Dependence between changes

If there are dependences between changes of action sets for different states, whether a certain action was available or not affects occurrence of changes of other actions which depend on that action. Once we know that a change C_1 has occurred then we can know whether a change C_2 that depends on change C_1 occurs or not with more confidence, i.e. with probability close to 1 or 0. Therefore, when availability of a certain action is known in an episode, we define augmented states x as states including the information. A binary variable f_i is defined which represents the availability of action i: if action i is unavailable, $f_i = 0$, and if action i is available, $f_i = 1$. A vector f is defined as the vector composed of all relevant f_i, and x is defined as $x = \begin{bmatrix} f \\ s \end{bmatrix}$.

Action values $\bar{Q}$ based on augmented states x are updated by

$$\begin{aligned} \bar{Q}(x_t, a_t) \leftarrow & (1-\alpha)\bar{Q}(x_t, a_t) \\ & +\alpha(1-P)\{r_{t+1} + \gamma \max_a \bar{Q}(x_{t+1}, a)\}. \end{aligned} \tag{4}$$

In each episode, at first, the agent selects actions based on Q. When availability of a certain action that has influences on changes of other actions is known, the agent selects actions based on $\bar{Q}$ until that episode ends. It means that the agent modifies the policy by using real time information.

4 SIMULATIONS

4.1 Problem Setting

We use an example road network shown in Fig.3, where the branches represent road sections and the nodes represent intersections.

In Fig.3, a number in a node is a state number, and a number at the side of a branch is reward obtained by selecting that branch as an action. A large reward is desirable. The starting state is state 0 and the termination state is state 22. $E_{i,j}$ represent an event that an action $a_{i,j}$ moving from node i to node j becomes unavailable.

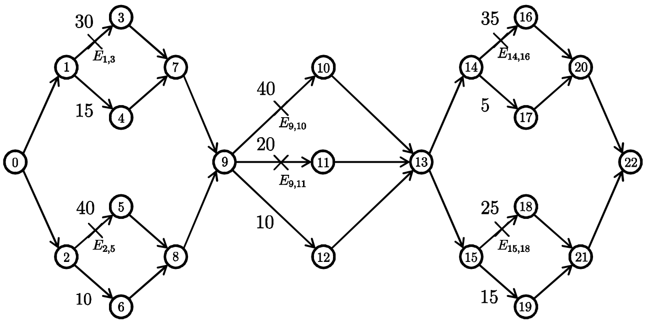

Fig. 3. An example road network.

In branches whose rewards are not explicitly shown, the rewards are all 0 to simplify the problem because those branches are not affected by changes.

Moreover, we assume dependences only between changes of actions selectable in an episode.

For example, we assume that there is dependence between $E_{1,3}$ and any changes except $E_{2,5}$ because the agent can select any changeable actions except $a_{2,5}$ after selecting $a_{1,3}$, but we assume no dependence between $E_{1,3}$ and $E_{2,5}$ since the agent cannot select $a_{2,5}$ after selecting $a_{1,3}$.

4.2 Parameter Setting

Occurrence probabilities of each change in Fig.3 are defined as shown in Table1.

Table 1. Occurrence probabilities.

	Conditional event E				
Probabilities	None	$E_{1,3}$	$E_{2,5}$	$E_{9,10}$	$E_{9,11}$
$P(E_{1,3} \mid E)$	0.35	-	-	-	-
$P(E_{2,5} \mid E)$	0.65	-	-	-	-
$P(E_{9,10} \mid E)$	0.65	0.8	0.2	-	-
$P(E_{9,11} \mid E)$	0.35	0.8	0.8	0.65	-
$P(E_{14,16} \mid E)$	0.65	0.8	0.5	0.2	0.5
$P(E_{15,18} \mid E)$	0.35	0.8	0.5	0.8	0.5

Conditional probabilities given two events, e.g. the conditional probability of $E_{i,j}$ given $E_{k,l}$ and $E_{m,n}$, are not shown in Table1, but we assume that all of these conditional probabilities are 0.5.

We do not treat conditional probabilities given three or more events, and this reason is explained later.

The agent only knows whether actions subject to changes are available or not after trying to take those actions. For example, the occurrence of event $E_{1,3}$ is known after the agent selects an action $a_{1,3}$.

Parameters used in simulations are shown below:

- the number of episode: $10,000,000$,
- the learning rate α: 0.1,
- the discount factor γ: 0.9,
- the method for action selection: $\epsilon-$greedy (ϵ is decreased from 0.5 to 0.05 step by step).

We perform simulations with different parameters in $F(p)$ namely DMs with different subjective interpretation as below:

DM(a) $\tau = 0.5, \sigma = 0.1$,
DM(b) $\tau = 0.3, \sigma = 0.1$,
DM(c) $\tau = 0.5, \sigma = 1$.

Parameter settings of DM(a) and DM(b) correspond to DMs considering that changes are likely to happen when the objective occurrence probabilities equal to or higher than 0.5 and 0.3, respectively because τ is the parameter adjusting threshold values, that is, DM(b) takes more risks than DM(a). DM(a) and DM(c) have the same τ value, but different σ values. DM(c) has gentler slope and therefore does not consider the change is likely to happen until its objective probability well exceeds 0.5.

4.3 Results

Optimal routes based on learned action value functions are shown as belows. ($E_{1,3}$) indicates that the shown route is the optimal route when the occurrence of $E_{1,3}$ is known by selecting $a_{1,3}$.

DM(a) $\tau = 0.5, \sigma = 0.1$
0-1-3-7-9-11-13-15-18-21-22
0-1-4-7-9-12-13-15-19-21-22 ($E_{1,3}$)
0-1-3-7-9-12-13-14-16-20-22 ($E_{9,11}$)
0-1-4-7-9-12-13-14-16-20-22 ($E_{1,3} \cap E_{9,10}$,$E_{1,3} \cap E_{9,11}$)

DM(b) $\tau = 0.3, \sigma = 0.1$
0-1-4-7-9-12-13-15-19-21-22
0-1-4-7-9-12-13-15-19-21-22 ($E_{1,3}, E_{9,11}$)
0-1-4-7-9-12-13-14-16-20-22 ($E_{9,10}$)
0-1-4-7-9-12-13-15-19-21-22 (($E_{1,3} \cap E_{9,10}$)$\cup$($E_{1,3} \cap E_{9,11}$))

DM(c) $\tau = 0.5, \sigma = 1$
0-1-4-7-9-12-13-15-18-21-22
0-1-4-7-9-12-13-15-19-21-22 ($E_{1,3}$)
0-1-4-7-9-12-13-14-16-20-22 ($E_{9,10}$,$E_{9,11}$)
0-1-4-7-9-12-13-14-16-20-22 ($E_{1,3} \cap E_{9,10}$,$E_{1,3} \cap E_{9,11}$)

It is found that DM(a) does not select a branch that becomes unavailable at the probability 0.65 but selects a branch that becomes unavailable at the probability 0.35. Note that unconditional occurrence probabilities are either 0.35 or 0.65 only. Moreover, if $E_{1,3}$ occurs, DM(a) selects a route without changes because all probabilities of changes increase to 0.8. This also shows that DM(a) selects the new optimal routes in response to real time information about the occurrence of changes.

DM(b) and (c) also select appropriate routes for their own subjective risk parameters τ and σ, that is DM(b) selects route in risk averting manor and DM(c) selects route paying less attention to the difference in objective probabilities as compared to DM(a), and DMs select better routes in response to real time information if changes of actions are caused on the way.

We do not treat conditional probabilities given three or more events as described in section 4.2. Simultaneous changes of three or more actions seldomly happen. Therefore, there are few chances that the agent selects actions when three or more events occur. In RL, because action values are updated only after taking those actions, the values of actions under the condition that three or more events occur do not converge. For this reason, we do not consider conditional probabilities given three or more events.

5 CONCLUSIONS

In this paper, we proposed RL approach to obtain the optimal policies in MSDM problems with changes of action sets considering DM's subjective views. It has been shown that the optimal policies are learned which correspond to DM's subjective views.

In contrast to standard RL methods that find the only one optimal policy, the proposed method can select routes in response to real time information during performing actions. This feature is promising in practical applications.

In future work, the issue must be solved that the conditional action values given three or more events do not converge. We have assumed that all probabilities are known, but it is desirable to estimate probabilities by learning.

REFERENCES

[1] Sutton.R.S and Barto.A.G: Reinforcement Learning - An Introduction, The MIT Press (1998).

[2] Peter Geibel and Fritz Wysotzki: Risk-Sensitive Reinforcement Learning Applied to Control under Constraints, Journal of Artificial Intelligence Research 24 81-108 (2005).

[3] Makoto Sato and Shigenobu Kobayashi: Variance-Penalized Reinforcement Learning for Risk-Averese Asset Allocation, Proc. of IDEAL 2000 244-249 (2000).

[4] Takeshi Shibuya: A study on reinforcement learning in unstationary dynamic environments, Proc. of SSI 2010 3B1-2 (2010) (in Japanes).