日本知能情報ファジィ学会支部研究部会合同研究会2010
－東北・関東支部＆知的制御・ヒューマンインタフェース研究部会－
（2010.7.2(金),3(土), 伝国の杜[山形県米沢市]）

A proposal of a Substitute Target Learning based Inverted Pendulum Swing-Up Control System

補助目標学習による倒立振り子の振り上げ制御を提案

○ Syafiq Fauzi Kamarulzaman　Seiji Yasunobu
Intelligent control system laboratory, University of Tsukuba

Abstract: 人間は制約に対してうまく行動を決定することができる。人間は体の状態を認識して、関連行動を選択することが可能と考えている。この行動を選択するとき、知識を用いて行動を選択する。この知識で最終目標を目指すために、補助目標を組み立てることができる。本研究では、以上の人間の能力を模倣、倒立振り子の振り上げ制御に適用した。振り子は補助目標の知識で振り上げて、最終目標という倒立状態を目指す。この行動を実現するために、強化学習の知識を用いて、適切な補助目標を生み出す。学習がうまく行えるように、振り子の行動はクラスターに分割して、そのクラスターの変化によって、適切な処理を選ぶ。最後に、シミュレーションで以上の内容の有効性を確かめた。

1 Introduction

Human have the ability to effectively execute any action under any constraints. Human can perform an effective continuous action by choosing suitable actions by recognizing the state of their body. In order to select these suitable actions, human use their acquired knowledge to select depending on the state of their body. If the final action acts as a final movement to achieve the final target, several substitute target must be performed before their body reach a state where a final action could be perform. Human knowledge plays a major role in selecting these substitute targets in order to make these continuous actions possible.

In this research, Substitute Target Knowledge is implemented to an Inverted Pendulum Swing-up control in order for it to imitate the above human ability. Reinforcement learning is used for the system to construct and rewrite its own knowledge which will be used to generate substitute targets. The pendulum movement state is separated into several clusters in order to make the evaluation of the substitute target easier. The pendulum downwards position act as initial state while the pendulum inverted state act as the final target. The objective of this system is to generate several substitute targets for the pendulum cart while swinging the pendulum towards its final state. Strong non-linearity in the pendulum system is the main reason an Inverted Pendulum System is used in this research.

2 Substitute Target Knowledge

Reinforcement learning is used in this research in order to implement a human-liked knowledge into the control system. The substitute target is constructed during certain state several times for the system to be able to propel the control object towards its final state. Thus, the distance between the control object current position and the substitute target is used in the reinforcement learning algorithm.

Q-learning algorithm is used in order to construct the substitute target knowledge. Instead of evaluating action, a, the system evaluate the distance to the substitute target, Δx in the value knowledge Q.

$$Q(s, \Delta x) \tag{1}$$

Therefore, the substitute target, xT is generated by summing the control object current position, x_{now} and the distance towards the substitute target, Δx.

$$xT = x_{now} + \Delta x \tag{2}$$

The control object parameter will be clarified as state, s which is use in equation (1). The substitute target distance, Δx is generated depending on the parameter state, s and selected among the index of Δx using *roulette* selection method. This event will occur in one of the control object movement state cluster.

3 Control object movement state clusterization

Movement state clusterization is a method used to separate the control object situation into several cluster. It is easier to determine the process needed to control the control object based on the pendulum situation. Therefore, in this case of Inverted Pendulum Control, the clusterization is made depending on the angular displacement of the pendulum, θ, and the angular displacement speed of the pendulum, ω. Since the control output is based on the substitute target, it is important to use the above parameters clusters to help the system achieve its target fast and accurate.

According to Figure 1, each cluster is given a cluster number for it to be recognized easier. The cluster number used to determine the previous cluster and the current cluster which is use to select the process suitable for the current pendulum situation. Therefore, the process needed by the system can be selected according to the changes between those two clusters.

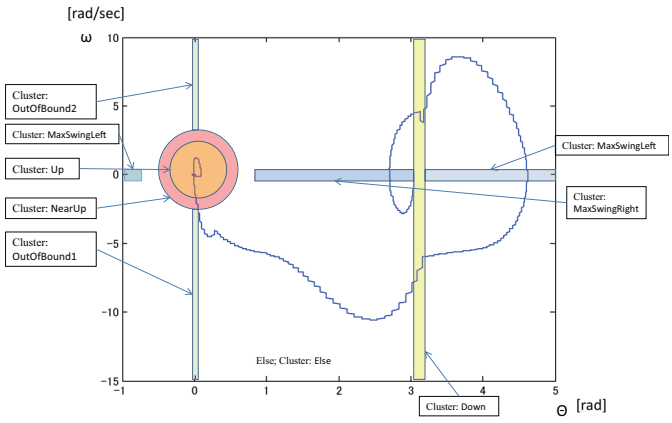


Fig. 1: Pendulum State Clusterization

4 Control System Design

The method used in section 2 and section 3 is combined to construct a working control system for the pendulum. The proposed system shown in Figure 2, consist of 3 major areas. These areas are the Control Area, Knowledge Learning Area, and Cluster Recognition area. The Control Area is used to change the output of substitute target distance into voltage output for the control object motor. The knowledge Learning Area is the area where reinforcement learning occurred. The reinforcement learning algorithm used in the learning area is Q-learning as shown in equation 3.

$$Q(s,\Delta x) = (1-\alpha)Q(s,\Delta x) + \alpha[r + \gamma \max_{\Delta x'} Q(s',\Delta x')] \quad (3)$$

r : Reward for the state.
α: Learning rate.
γ: Discount rate

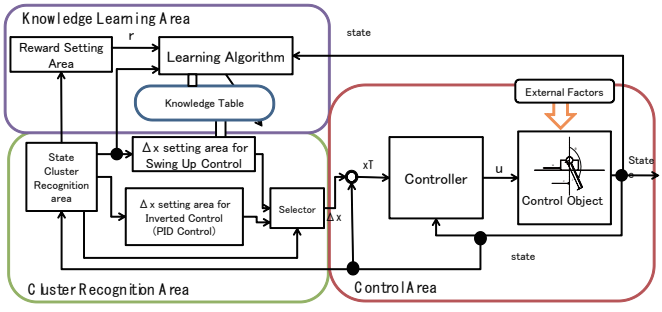
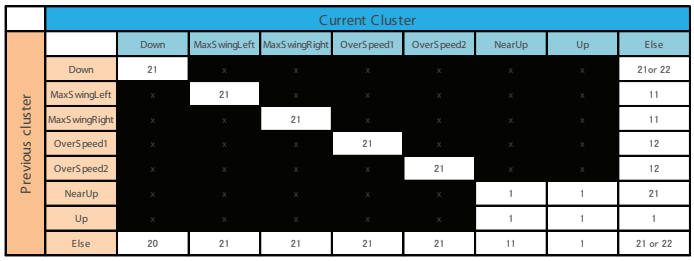


Fig. 2: The proposed system

The final area is the Cluster Recognition Area which will recognize the current cluster based on the current control object parameters. This area plays a major role on selecting a suitable process for the control object. The suitable processes are selected according to the situation, either to generate reward, rewrite the knowledge table, or to generate substitute target for the Swing-up control and Inverted Control.

4.1 Process selection via cluster change

The cluster recognize in the Cluster Recognition Area will be recognize according to Figure 1. The previous cluster will be recorded within the system and it will be compared to the current cluster in order to select a suitable process. The process will be selected according to Figure 3.



Process: 1 Stabilization Controller & Reward Setting: Reward = r
21 none
10 Reward Setting : Reward = 0
11 Reward Setting : Reward = r-Δr (Δr : discount rate)
12 Reward Setting : Reward =-r
20 Swing Up Controller & Table Update

Fig. 3: Process Selection

By using this method in selecting process, the system will supply clear view on processes run according to the situation cluster. This is also used to effectively configure necessary rewards for the reinforcement learning algorithm depending on the situation. Thus, it helps the system to effectively configure each substitute target needed to control the control object upon achieving its control objective.

5 Simulation

The simulation for the system is done on an Inverted Pendulum which used as a control object as shown in Figure 4. The simulation is run for 2000 episodes where each episode consists of 10 second run time. The detail for the pendulum used in this simulation is as shown in Table 1.

Table 1: Simulation device's parameters setting

cart mass	3.117 (kg)
Pendulum Length	0.2 (m)
Pendulum Mass	0.08 (kg)
Pendulum Inertia	$\begin{bmatrix} 0.0012 & 0 & 0 \\ 0 & 1.6\times10^{-7} & 0 \\ 0 & 0 & 0.0012 \end{bmatrix}$ (kgm^2)

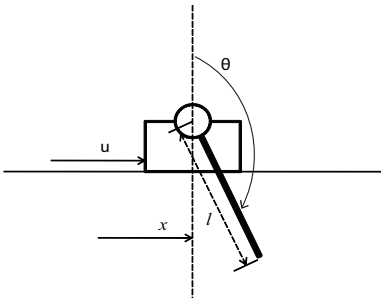


Fig. 4: Inverted pendulum diagram

In order to assure the learning process proceeds perfectly, the most adequate array of knowledge is previously constructed. This is made to ensure that the program could sustain a perfect knowledge without being interrupted by any unknown programming bug. Therefore an almost perfect knowledge is inserted and the learning process is viewed during the simulation. The system will evaluate the most suitable substitute target which is based on the carts movement. In this case, the substitute target is the sum of the carts movement and the substitute target distance supplied from the knowledge table.

The knowledge table used in this simulation is a 3 dimensional array. The knowledge table state, s is a combination of 2 parameters which is the pendulum cart position, x and the pendulum angular displacement velocity, ω. The 3 dimensional array need to be converted into a 2 dimensional array in order to obtain an adequate result for analysis. Therefore, parameters x and ω are combined and numbered as shown in a table in Figure 4. The table in Figure 4 is constructed using Table 2 state parameters.

	Angular Displacement Speed, ω							
Cart Position, x	ω_1	ω_2	ω_3	...	...	...	...	ω_{29}
x_1	1	22	43	...	...	...	...	588
x_2	2	:	:	...	...	...	...	589
x_3	:	:	:	...	...	...	....	590
:	:	:	:	...	...	...	...	:
:	:	41	62	...	...	...	...	608
x_{21}	21	42	63	...	...	...	...	609

Fig. 5: Knowledge Parameters States combination numbering

The Q-learning parameters in Table 2 are used in the section 2 systems learning algorithm. These parameters are previously selected adequately.

5.1 Simulation Result

From the result of Figure 8 and 9, it is understood that the learning occurred smoothly using the previously constructed knowledge without inadequately increases or decreases its initial knowledge as shown if Figure 6 and 7. This shows that the program is able to write and rewrite the knowledge correctly. However, from several trials, it is understood that the program still have several bug in applying rewards for the knowledge to be rewritten. This is because it is possible for the system not only to manipulate angular velocity acceleration but also deceleration in order to select a suitable substitute target.

Table 2: Q-learning parameters

Discount rate,α	0.5
Learning rate,γ	0.2

State Parameters	Range	Intervals
Cart Position, x	-1.0 $\sim$ 1.0 (m)	0.1(m)
Pendulum angular velocity, ω	-14 $\sim$ 14 (rad/s)	1 (rad/s)

action parameters	Range	Intervals
Substitute target distance, Δx	-0.2 $\sim$ 0.2 (m)	0.05(m)

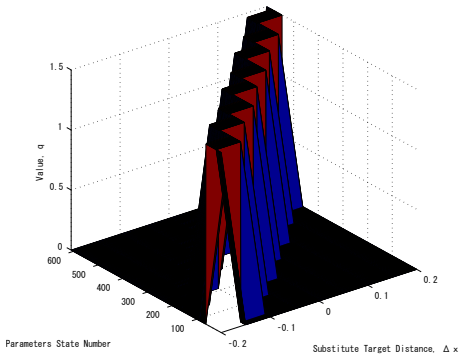


Fig. 6: Knowledge state and substitute target distance evaluation Initial Surface

In Figure 10, it is understood that the pendulum are able to move similar to result expected as in Figure 1. This concludes that the clusterization of the pendulum movement did contribute in producing a successful swing up result.

From Figure 11, it is able to see that the cart position changes according to substitute target distance. Substitute target distances is generated until up to 6 second when the stabilization control occured. Therefore, the swing-up process is assume to have produced an expected result.

The simulation is run from an adequate intial knowledge in order to reduce the influence of any unwanted bug in the program. A simulation with random initial knowledge is scheduled to be done in the future.

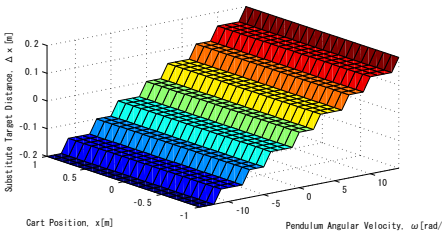


Fig. 7: Best Substitute Target Distance based on Knowledge State Initial Surface

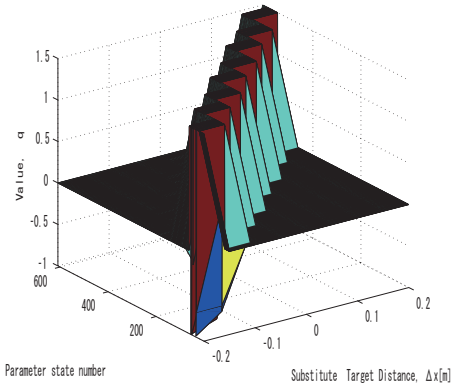


Fig. 8: Knowledge state and substitute target distance evaluation result

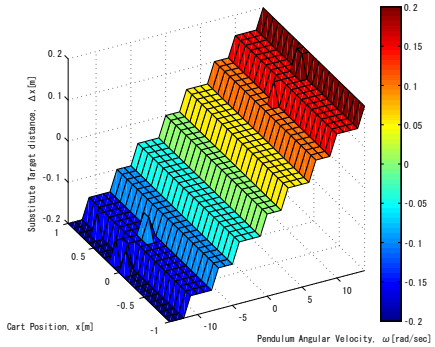


Fig. 9: Best Substitute Target Distance based on Knowledge State

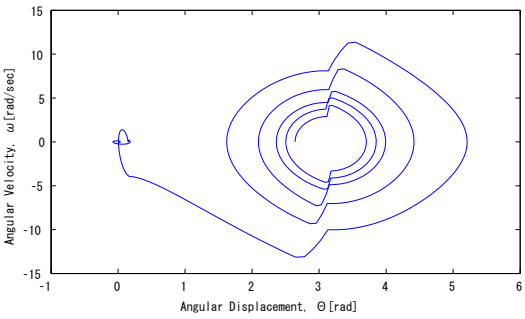


Fig. 10: Angular displacement and angular velocity relation from a succesful swing up result

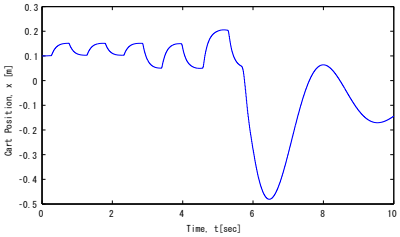


Fig. 11: The cart movement of a succesful swing up result

6 Conclusion

In this research, Substitute Target Knowledge is implemented to an Inverted Pendulum Swing-up control in order for it to imitate the above human ability. Reinforcement learning is used for the system to construct and rewrite its own knowledge which will be used to generate substitute targets. It is understood that by seperating the pendulum movement state into several clusters, the evaluation of the substitute target can still occur effectively. The objective of this system which is to generate several substitute targets for the pendulum cart while swinging the pendulum towards its final state, had been accomplished by using a nearly adequate knowledge as its initial state. Although the simulation have not been done using other initial knowledge, it is known that reinforcement learning is still occured in this system. Thus it strongly shows that the possibility of the system to learn using other initial knowledge is high. Therefore, a system which make use substitute target knowledge and control object state recognition are capable to propel a control object towards its final target from its initial condition by using multiple substitute targets.

References

[1] Richard S. Sutton, Andrew G. Barto, "Reinforcement Learning An Introduction ", MIT Press(1998).

[2] H. Iima and Y. Kuroe, "Swarm Reinforcement Learning Algorithms Based on Actor-Critic Methods ", 35th SICE Symposium on Intelligent Systems , pp.9-14, 2008.

[3] E. Kawana and S. Yasunobu, "An Intelligent Control System Using Object Model by Real-Time Learning ", SICE Annual Conference , pp.2792-2797, 2007.

[4] T. Matsubara and S. Yasunobu, "An Intelligent Control Based on Fuzzy Target and Its Application to car like Vehicle ", SICE Annual Conference , 2004.