A learning-based control system by knowledge acquisition within constrained environment

*Syafiq Fauzi Kamarulzaman
Intelligent Interaction Technologies
University of Tsukuba
syafiq.ics@edu.esys.tsukuba.ac.jp

Takeshi Shibuya
Intelligent Interaction Technologies
University of Tsukuba
shibuya@iit.tsukuba.ac.jp

Seiji Yasunobu
Intelligent Interaction Technologies
University of Tsukuba
yasunobu@iit.tsukuba.ac.jp

Abstract: Knowledge acquisition is important in order to build a knowledge based system. One of the methods used in acquiring knowledge is reinforcement learning. Reinforcement learning is commonly defined as a try and error style learning that occurred in episodes. This is difficult to ensure a real control object safety condition since a control object is restricted to its environmental constraints. Therefore, a control system which can acquire knowledge within constrained environment, and use it to achieve a control objective in continuous time is proposed. This control system is applied on an inverted pendulum control system and its effectiveness is confirmed through a series of simulations.

1 INTRODUCTION

A knowledge-based system is a system equipped with knowledge based on a human knowledge. Some knowledge is based on human expertise in specific objectives, applied into a system which is known as an expert system[1]. However, some knowledge needed to be constructed when human expertise knowledge is either not available or needed to be updated based on changes of constraints in environment.

One of the methods used for constructing knowledge is reinforcement learning [2],[3],[4]. Since reinforcement learning is a try and error style learning process, real control object could be damaged during the learning process [5]. This is because rough shock and vibration could be generated during constraints encounter. Therefore, it is difficult to have the entire learning process conducted on a real control object. It is important for the control system to learn to consider environmental constraints during learning since shock and vibration could alter the real control object physical properties, which could lead to failure on achieving a control objective.

Conventional control uses various methods on applying reinforcement learning into a control system [6],[7],[8],[9]. One of the method proposed could control a control object by using substitute targets based on learned knowledge [6]. Substitute targets are targets arranged to help the control object achieve its target state which is the control objective. From this conventional method, it is proven that it is possible to create substitute target knowledge and control using the knowledge in order to achieve a control objective. However, the conventional method was only applied in simulation using episodic sequences without focusing on considering constraints encounter influences.

In this research, a different approach based on the conventional method is proposed in order for the system to be able to acquire knowledge to achieve its control objective within assigned constraints. This system is designed for real control object application therefore a method applicable on real machine is arranged. Constraints knowledge is added to make the system capable to acquire knowledge in order to achieve its control objective in assigned constraints within controllable parameters. This system is applied on an inverted pendulum control system and tested in simulations assigned with different sets of selected constraints.

2 CONTROL SYSTEM DESIGN

In this research, the proposed system is designed to be applicable on an inverted pendulum system. The proposed system for this research is shown in Fig.1. This system is constructed based on a system proposed in our previous research [6]. Based on this structure, there are three major areas for specific purposes. These areas are the (i) Control Area, (ii) Recognition Area, and (iii) Learning Area.

The pendulum system used in this research consists of a cart and a pendulum as shown in Fig. 2. Force is applied to the cart moving along a track as to swing the pendulum hinged to the cart towards its inverted state. A success is said to occur if the pendulum achieved the inverted state. The pendulum and the cart will be returned to its initial position for another swing up attempt after being in inverted state for a few seconds.

2.1 Control Area

There are three major sections that control the cart movement in the Control Area. These sections are responsible on
setting substitute target, which will be based on states. These sections are swing up control section, inverted control section and initialization control section.

The swing up control section controls when the pendulum is in downwards position as in Fig. 2 (b). In the swing up control section, substitute target displacement, $\Delta x$ is selected from the substitute target knowledge. Substitute target is constructed during the pendulum downwards position based on the substitute target displacement, $\Delta x$, provided by the substitute target knowledge. Substitute target, $x_T$ is calculated by adding the selected substitute target displacement, $\Delta x$ to the current cart position, $x_{\text{now}}$ as

$$x_T = x_{\text{now}} + \Delta x. \quad (1)$$

The inverted control section also controls using substitute target, $x_T$ calculated as in (1). This control section controls when the pendulum is near to its inverted state as shown in Fig. 2(a). However, substitute target displacement, $\Delta x$ is generated through PD control.

The initialization section generates the control commands to move the cart towards the initial position. This occurs after the pendulum achieves the inverted state as shown in Fig. 2(a) or after constraints encounter. This section also controls using substitute target, $x_T$ calculated based on substitute target displacement, $\Delta x$ generated through PD control as the inverted control section.

2.2 Recognition Area

The Recognition Area is the area where the control object state is managed into sets of state clusters. This area also acts as a switch which functions based on state clusters. This helps the Control Area to select a suitable control sections output. The state clusters created is also needed to determine rewards based on current state in the Learning Area. In this paper, the pendulum cart position, $x$, pendulum angle, $\theta$, and pendulum angular velocity, $\omega$ are the clustered state parameters.

Constraints information provided by the assigned constraints knowledge will also be included in these clusters. In case of this research control object, which consists of the cart and the pendulum, the constraints can be divided into two as shown in Fig. 3. Constraints shown in Fig. 3 (a), are constraints restricting the cart movement when swinging the pendulum towards its inverted state. Constraints shown in Fig. 3(b) are constraints restricting the pendulum movement when the cart attempt to swing it towards inverted state. However in this paper, we only focused on controlling the control object while concerning constraints as shown in Fig. 3(a).

2.3 Learning Area

The Learning Area is the area where the substitute target knowledge is rewritten based on state clusters. In this paper, the knowledge is rewritten when the pendulum is in full downwards state as shown in Fig. 2 (b) before the Control Area selects the substitute target displacement, $\Delta x$ which is needed to move the cart. The substitute target knowledge, $Q$ is a value function which is rewritten using Q-learning algorithm [2], as in

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

where state, $s$ used in this paper is the cart position, $x$ and the pendulum angular velocity, $\omega$.

Reward is important to the system since reward defines the goal in reinforcement learning. Rewards helps determining substitute target displacement, $\Delta x$ in a certain states that has the highest value in the value function. Reward, $r$ in (2) is determined by state cluster numbers at a precise moment. In this paper, rewards were assigned based on the control object current state clusters as shown in TABLE I.

3 SIMULATION EXPERIMENT AND RESULTS

3.1 Simulation experiment setup

The simulation is about to have the control object to be able to learn to achieve the control objective within the assigned constraints. The simulation for our proposed system was conducted in MATLAB Simulink where the parameters used are based on a real control object parameters from an experiment device as shown in Fig. 4.

In order to make the system applicable on a real control object, we had specified several required behavior for the system control object, which is:
If 3000 second run time ends, stop simulation.
If pendulum is in downwards position, initialize after 10 times knowledge renewal.
If the pendulum is in inverted position, initialize after 3 second.
If cart position is between constraints, initialize after constraints encounter.
Pendulum swing attenuation during initialization.

The Q-learning parameters in TABLE II were used in the proposed system’s Q-learning algorithm. These parameters were pre-experimentally selected. Fig. 5 shows the substitute target knowledge figure at the beginning of the simulation. This figure is constructed based on an analyzed knowledge which was only based on pendulum swing-up control without considering the cart position.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Range</th>
<th>Intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cart Position, x[m]</td>
<td>-1.0 ~ 1.0</td>
<td>0.2</td>
</tr>
<tr>
<td>Pendulum Angular Velocity, ω [rad/s]</td>
<td>-14 ~ 14</td>
<td>2</td>
</tr>
<tr>
<td>Substitute Target Displacement, Δx [m]</td>
<td>-0.2 ~ 0.2</td>
<td>0.05</td>
</tr>
</tbody>
</table>

TABLE III: INITIAL STATE AND TARGET STATE

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Initial State</th>
<th>Target State</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cart Position, x[m]</td>
<td>0</td>
<td>xT</td>
</tr>
<tr>
<td>Pendulum Angle, θ [rad]</td>
<td>π</td>
<td>0</td>
</tr>
<tr>
<td>Cart Velocity, v [m/s]</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pendulum Angular Velocity, ω [rad/sec]</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The control object initial and target state is shown in TABLE III. The initial state, based on Fig. 6, is the state when the pendulum is in downwards position as in Fig. 2 (b), where the pendulum angle, θ = π [rad]. The cart position, x is in the middle of the track, x = 0[m], while the cart velocity, v and pendulum angular velocity, ω is zero. The target state is the control objective state as show in Fig. 2(a), where the pendulum angle, θ is 0 [rad]. The target state cart position is the final substitute target, xT selected during the last swing up process. After the inverted state is achieved, the new initial cart position, x is the last inverted state cart position. This is to make sure that the system will not repeat the same swing-up strategy as the previous attempt trial thus, will increase the amount of updated area within knowledge.

In this paper, three simulations was conducted by only concerning cart movement, x constraints as shown in Fig. 3(a). These cart movement constraints are selected pre-experimentally. The diagram concerning the cart movement constraints is shown in Fig. 7. The cart movement, x constraints shown in Fig. 7 are:
(a) Case 1(P8M8), −0.8 < x < 0.8 : Both constraint in left and right side
(b) Case 2(P4M4), −0.2 < x < 0.6 : Bigger constraint in both left and right side

![Fig. 4: The experiment device based for the simulation (Japan E.M. Co.,Ltd.)](image)

Fig. 4: The experiment device based for the simulation (Japan E.M. Co.,Ltd.)

![Fig. 5: The initial substitute target knowledge.](image)

Fig. 5: The initial substitute target knowledge.

![Fig. 6: The control object diagram](image)

Fig. 6: The control object diagram
3.2 Simulation Result

Simulation results include the system substitute target knowledge, $\Delta x$ after 3000 seconds simulation run time. The after simulation figure of substitute target knowledge, $Q(s, \Delta x)$ is the knowledge structure after improvement during simulations. During the simulation, the system will configure its own swing up strategy. The number of trial conducted depends on initialization frequency since the whole simulation run time includes swing-up control, initial control and initialization control.

3.2.1 Simulation result for P8M8.

In this simulation, the cart position constraints exist in both left and right side as shown in Fig. 7(a). During this simulation, there was 427 swing attempt trials conducted within 3000 seconds. Successful and failed trial can be understood from Fig. 8. There is more failed trial in the beginning of the simulation but successful trial increased over time. The substitute target knowledge, $Q(s, \Delta x)$ at the end of the simulation is shown in Fig. 9.

Comparing Fig. 9 and Fig. 5, the knowledge within the constrained area was changed. This is because only the knowledge within the constrained area was rewritten during the simulation.

3.2.2 Simulation result for P4M4.

In this simulation, the cart position constraints were assigned heavily on the track left side and slightly on the track right side as shown in Fig. 7(b). During this simulation, there was 588 swing attempt trials conducted within 3000 seconds. Successful and failed trial can be understood from Fig. 10. There is more failed trial in the beginning of the simulation but successful trial increased over time. However, failed trial is becoming frequent for several trials before successful trial frequency increased. The substitute target knowledge, $Q(s, \Delta x)$ at the end of the simulation is shown in Fig. 11.

Comparing Fig. 11 and Fig. 5, again, the knowledge within the constrained area was changed.
3.2.3 Simulation result for P2M6.

In this simulation, the cart position constraints were assigned heavily on the track right side while slightly on the left side as shown in Fig. 7(c). During this simulation, there was 462 swing attempt trials conducted within 3000 seconds. Successful and failed trial can be understood from Fig. 12. There is more failed trial in the beginning of the simulation but successful trial increased over time. However, successful trial frequent is lesser compare to the simulation using constraints shown in Fig. 7 (a). The substitute target knowledge, \( Q(s; \Delta x) \) at the end of the simulation is shown in Fig. 13.

Comparing Fig. 13 and Fig. 5, as the other simulation results, only the knowledge within the constrained area was changed.

3.3 Discussion

Based on the results, a successful swing up process did not occur continuously between trials. This is because the pendulum did not able to achieve the control objective using the same swing up strategy since the initial cart position, \( x \) is change to be the last inverted state cart position. However, by this method, more area in substitute target knowledge, \( Q(s; \Delta x) \) is explored. Result from the three simulations each shows that only the knowledge within the constrained area were used and rewritten by the system. From this result, it is understood that the proposed system is able to acquire knowledge to achieve its control objective within assigned constraints.

Fig. 14 shows the cart movement when the cart position, \( x \) is constrained to \((-0.8 < x < 0.8)\). Based on Fig.14, the
the cart position, inverted state where inverted control occurs. Fig. 15 shows cart moves to swing the pendulum and succeed on achieving inverted state around the same time as Fig. 15. Fig. 14, 15 and 16 all shows that the system are able to control within assigned constraints using the acquired knowledge. Therefore, the system is known to be able to achieved its control objective while in constrained environment by using the acquired knowledge.

4 CONCLUSION

In this research, a system which is able to acquire knowledge to achieve its control objective within assigned constraints is proposed. This method could control a control object by using substitute targets provided by substitute target knowledge. Reinforcement learning is used in the system to change or renew its own knowledge which will be used to generate substitute targets. This system is designed for real control object application therefore a method applicable on real machine was arranged and conducted in simulation. Constraints knowledge was added to make the system capable to acquire knowledge in order to achieve its control objective in assigned environmental constraints.

The proposed control system were able to learn to consider environmental constraints during learning, therefore is assume to be able avoid shock and vibration that could lead to failure on completing a control objective. Although the proposed system was only assigned with the cart movement constraints, substitute target knowledge were rewritten as expected. In conclusion, a learning-based system by knowledge acquisition within constraints is achieved.

REFERENCES