2008 International Symposium on Nonlinear Theory and its Applications
NOLTA'08, Budapest, Hungary, September 7-10, 2008

Reinforcement Learning of Optimal Supervisor for Timed Discrete Event Systems

Tatsushi Yamasaki† and Toshimitsu Ushio‡

†Faculty of Engineering, Setsunan University
17-8 Ikeda-nakamachi, Neyagawa, Osaka, 572-8508 Japan
‡Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama, Toyonaka, Osaka, 560-8531 Japan
Email: yamasaki@ise.setsunan.ac.jp, ushio@sys.es.osaka-u.ac.jp

Abstract—Timed discrete event systems are a class of discrete event systems which can represent information of time of event occurrences. In this paper, we propose a synthesis method of a supervisor for such systems. The supervisor is constructed by reinforcement learning under the framework of the supervisory control proposed by Brandin and Wonham, and is optimal with respect to a language measure which is introduced by Wang and Ray. The proposed method is applicable to the situation which precise information of the system is unknown.

1. Introduction

Discrete event systems (DESs) are widely found in many artificial systems[1]. The supervisory control initiated by Ramadge and Wonham is a logical control method for DESs[2]. In the Ramadge and Wonham framework, a controller, called a supervisor, assigns a set of events permitted to occur for satisfying the control specifications. In some systems such as real time systems, some events are required to occur within designated time bounds. Timed DESs are a class of DESs which can represent such specifications. Brandin and Wonham proposed a supervisory control for timed DESs[3]. They introduced a tick event and forcible events to represent and control timed DESs.

In this paper, we propose a synthesis method of the supervisor for timed DESs. The proposed method is an extension of our previous works to timed DESs [6, 7, 8]. The method uses reinforcement learning to learn the supervisor and is optimal with regard to a language measure. The language measure is a performance index for the languages generated by DESs[4, 5]. So, it is possible to evaluate the performance of the system quantitatively. An optimal scheduling in soft-real time systems using language measure is also proposed[9]. By using reinforcement learning, the supervisor learns the control patterns based on rewards. Therefore the proposed method is applicable under the imprecise description of specifications and uncertain environment.

This paper is organized as follows. The model of the timed DES and the language measure are introduced in section 2. Description of the system by the Bellman equations is shown in section 3. The learning method of the optimal supervisor is proposed in section 4. Conclusions are presented in section 5.

2. Preliminaries

In this paper, we use the timed DES proposed by Brandin and Wonham[3]. The timed DES G is modeled by a 5-tuple $G = (X, \Sigma, \delta, x_1, X_m)$ where, $X = \{x_1, x_2, \cdots, x_n\}$ is a set of finite states, $\Sigma = \{\sigma_1, \sigma_2, \cdots, \sigma_m\}$ is a set of finite events, $\delta : X \times \Sigma \to X$ is a state transition function, $x_1 \in X$ is an initial state, $X_m \subseteq X$ is a set of marked states. Σ^* is a set of all finite strings over Σ including the empty string ϵ. The transition function is extended to the function $\delta : X \times \Sigma^* \to X$ in the ordinary way. Denoted by σ_i^k is the index set of events by which a transition from state x_i to x_k occurs, *i.e.*, $\sigma_i^k = \{ j \mid \delta(x_i, \sigma_j) = x_k \}$. Denoted by $\hat{\sigma}_i$ is the index set of active events at state x_i, *i.e.*, $\hat{\sigma}_i = \{ j \mid \delta(x_i, \sigma_j)$ is defined$\}$. The language $L(G, x_i)$ generated by the DES G starting from the state $x_i \in X$ is defined by

$$L(G, x_i) = \{s \in \Sigma^* \mid \delta(x_i, s) \in X\}. \quad (1)$$

Σ is partitioned into a set of normal events Σ_{act} and a special event $\sigma_m = tick$. The event *tick* represents "tick of the global clock". Σ_{act} is also partitioned into a set of controllable events Σ_c and a set of uncontrollable events Σ_{uc}.

In the original supervisory control[2], the supervisor controls the occurrence of controllable events so as to satisfy logical control specifications of the DES. For the timed DES, a set of forcible events $\Sigma_f \subseteq \Sigma_{act}$ is introduced. Denoted by $\hat{\sigma}_i^f$ is the index set of forcible events at state x_i, *i.e.*, $\hat{\sigma}_i^f = \{ j \mid \sigma_j$ is defined and $\sigma_j \in \Sigma_f \}$. The supervisor can force the occurrence of the forcible events. If a forcible event is forced, a tick of the clock is preempted by the event. In other words, the occurrence of tick is prohibited by forcing a forcible event, and the supervisor can control the DES to satisfy time constraint. Note that forcible events may either be controllable events or be uncontrollable events, and one of other permitted events may occur in the DES even if there are forced events.

For the purpose of evaluation of the DES G controlled by the supervisor S, a signed real measure called a language measure is introduced[4, 5]. The set of marked state X_m is partitioned into a set of desired states X_m^+ and that of undesired states X_m^-. The characteristic function $y : X \to [-1\ 1] \cup \{-\infty\}$ is defined as follows:

$$y(x_i) = y_i \in \begin{cases} [\,-1,\ 0\,) \cup \{-\infty\} & \text{if } x_i \in X_m^-, \\ \{\,0\,\} & \text{if } x_i \notin X_m, \\ (\,0,\ 1\,] & \text{if } x_i \in X_m^+. \end{cases} \tag{2}$$

y_i shows the evaluation of each state, and the fatal states for the DES are expressed by a special value $-\infty$. The rule of calculation for y_i and $-\infty$ is defined as follows:

$$\begin{aligned} -\infty < y_i \quad & \text{for any } y_i \in [-1, 1], \\ y_i + (-\infty) = -\infty \quad & \text{for any } y_i \in [-1, 1] \cup \{-\infty\} \text{ and,} \\ y_i \times (-\infty) = -\infty \quad & \text{for any } y_i \in (0, 1]. \end{aligned}$$

$Y = [y_1\ y_2\ \cdots\ y_n]^T$ is called a state weighting vector.

A supervisor S assigns disabling events and forcing events to the DES G. Denoted by d_i^S is the index set of disabling events at state x_i, *i.e.*, $d_i^S = \{\, j \,|\, \sigma_j \text{ is disabled at state } x_i \,\}$, and is called a disabling pattern. Denoted by f_i^S is the index set of forced events at state x_i, *i.e.*, $f_i^S = \{\, j \,|\, \sigma_j \text{ is forced at state } x_i \,\}$ and is called a forcing pattern. A pair of a disabling pattern and a forcing pattern is called a control pattern. $c_i^S = (d_i^S, f_i^S)$ denotes a control pattern at state x_i. Denoted by $\hat{\sigma}_i^S$ is the index set of active events at state x_i under the control of the supervisor S, and is defined as follows:

$$\hat{\sigma}_i^S = \begin{cases} \{\, j \,|\, \hat{\sigma}_i - d_i^S - \{m\}\} & \text{if } f_i^S \neq \emptyset, \\ \{\, j \,|\, \hat{\sigma}_i - d_i^S\} & \text{if } f_i^S = \emptyset. \end{cases} \tag{3}$$

Note that *tick* $(= \sigma_m)$ is preempted by forced events in the control pattern. Forced events are always permitted to occur, *i.e.*, $f_i^S \cap d_i^S = \emptyset$.

Denoted by $p_{ij}(c_i^S) \in [0, 1]$ is the occurrence probability of the event σ_j at state x_i under the control pattern c_i^S. A state transition cost $\pi_{ik}^S \in [0, 1]$ of the controlled system S/G is defined as follows:

$$\pi_{ik}^S(c_i^S) = \begin{cases} \displaystyle\sum_{j \in \hat{\sigma}_i^S} p_{ij}(c_i^S) & \text{if } \hat{\sigma}_i^S \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Π^S denotes a state transition cost matrix whose (i, k)-th element is π_{ik}^S.

Let $\xi_i^S = \xi(x_i, c_i^S)$ be a cost by disabling or forcing events at state x_i. $\xi^S = [\,\xi_1^S,\ \xi_2^S,\ \ldots,\ \xi_n^S\,]^T$ is called a controlling cost characteristic vector of a supervisor S.

Then, a performance vector μ^S of the controlled system S/G is given as follows[10, 11]:

$$\begin{aligned} \mu^S(\theta) &= [\mu_1^S(\theta)\ \mu_2^S(\theta)\ \cdots\ \mu_n^S(\theta)]^T \\ &= \theta[I - (1-\theta)\Pi^S]^{-1} Y^S, \end{aligned} \tag{5}$$

where $\theta \in (0, 1)$ is a constant parameter, $Y^S = Y - \xi^S$ is a modified characteristic vector, and $\mu_i^S(\theta)$ is a language measure of $L(S/G, x_i)$. It represents a quantitative performance index of the timed DES controlled by the supervisor.

3. Description by Bellman equations

We model the controlled system G/S by the Bellman equation. The supervisor S assigns a control pattern c_i^S at state x_i to the timed DES G. Therefore, the following Bellman equation holds[7]:

$$V^S(x_i) = \sum_{x_k \in X} \left\{ P(x_i, c_i^S, x_k) \left(r^*(x_i, c_i^S, x_k) + \gamma V^S(x_k) \right) \right\}, \tag{6}$$

where $V^S(x_i)$ is a discounted expected total reward at state x_i under the control by the supervisor S and called a value function, $P(x_i, c_i^S, x_k)$ is a probability of a transition from state x_i to x_k when a supervisor S assigns a control pattern c_i^S, $r^*(x_i, c_i^S, x_k)$ is an expected reward when a state transition from x_i to x_k occurs by assigning the control pattern c_i^S, and γ is a discount rate of reward.

In the DES G, an event occurs based on the given control pattern c_i^S. Therefore, the following equation holds:

$$P(x_i, c_i^S, x_k) = \sum_{j \in \hat{\sigma}_i^S} P_1(x_i, c_i^S, \sigma_j) P_2(x_i, \sigma_j, x_k), \tag{7}$$

where $P_1(x_i, c_i^S, \sigma_j)$ is a probability that an event σ_j occurs in the DES G when the supervisor S assigns the control pattern c_i^S at state x_i, and $P_2(x_i, \sigma_j, x_k)$ is a probability that the DES G makes a transition from the state x_i to x_k when an event σ_j occurs.

We assume that $P_1(x_i, c_i^S, \sigma_j)$ is determined as follows:

$$P_1(x_i, c_i^S, \sigma_j) = \begin{cases} \dfrac{p_f^*(x_i, \sigma_j)}{\displaystyle\sum_{k \in f_i^S} p_f^*(x_i, \sigma_k) + \sum_{k \in \{\hat{\sigma}_i - d_i^S - f_i^S - \{m\}\}} p^*(x_i, \sigma_k)} \\ \qquad \text{if } f_i^S \neq \emptyset \text{ and } j \in \hat{\sigma}_i^f, \\ \dfrac{p^*(x_i, \sigma_j)}{\displaystyle\sum_{k \in f_i^S} p_f^*(x_i, \sigma_k) + \sum_{k \in \{\hat{\sigma}_i - d_i^S - f_i^S - \{m\}\}} p^*(x_i, \sigma_k)} \\ \qquad \text{if } f_i^S \neq \emptyset \text{ and } j \notin \hat{\sigma}_i^f, \\ \dfrac{p^*(x_i, \sigma_j)}{\displaystyle\sum_{k \in \hat{\sigma}_i - d_i^S} p^*(x_i, \sigma_k)} \\ \qquad \text{if } f_i^S = \emptyset \text{ and } \hat{\sigma}_i - d_i^S \neq \emptyset, \\ 0 \\ \qquad \text{otherwise,} \end{cases} \tag{8}$$

where $p^*(x_i, \sigma_j) \in [0, 1]$ represents an weight of occurrence of the event σ_j at state x_i, and $p_f^*(x_i, \sigma_j) \in [0, 1]$ represents an weight of occurrence of the forced event σ_j at state x_i. Therefore the weight of occurrence of the event

changes if the event is forced. The weight parameters have the constraint as follows: For all state x_i

$$\sum_{k\in\hat{\sigma}_i^f} p_f^*(x_i,\sigma_k) + \sum_{k\in\hat{\sigma}_i} p^*(x_i,\sigma_k) = 1. \tag{9}$$

This means the sum of all weight parameters at each state is always 1.

Moreover, we assume a discount rate γ is determined as follows:

$$\gamma = \gamma(x_i, c_i^S, \sigma_j) = \begin{cases} 1-\theta & \text{if } j \in \hat{\sigma}_i^S, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Then we define a function $\pi^{*S} : X \times X \to [0,1]$ as follows:

$$\pi_{ik}^{*S} = \begin{cases} \displaystyle\sum_{j\in\hat{\sigma}_i^S} P_1(x_i, c_i^S, \sigma_j) & \text{if } \hat{\sigma}_i^S \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Let Π^{*S} be a matrix whose (i,k)-th element is π_{ik}^{*S}.

We define the reward $r^*(x_i, c_i^S, x_k)$ as follows:

$$r^*(x_i, c_i^S, x_k) = r^*(x_i, c_i^S) = y(x_i) - \xi_i^S(x_i, c_i^S), \tag{12}$$

which means the reward is based on the evaluation of the current state x_i and the cost of the assigned control pattern c_i^S.

By using the above definitions and assumptions, if a state transition is deterministic, Eq. (6) is transformed into

$$V^S(x_i) = r^*(x_i, c_i^S) + (1-\theta)\sum_{x_k\in X}\left(\pi_{ik}^{*S} V^S(x_k)\right). \tag{13}$$

We define a vector V and R as $V = [V^S(x_1)\cdots V^S(x_n)]^T$ and $R = \left[r^*(x_1, c_1^S)\cdots r^*(x_n, c_n^S)\right]^T$ respectively. Then the following equation holds:

$$V = R + (1-\theta)\Pi^{*S} V, \tag{14}$$

$$V = [I - (1-\theta)\Pi^{*S})]^{-1} R. \tag{15}$$

By comparing Eq.(5) and Eq(15), it shows that the value function vector V and the performance vector $\mu^S(\theta)$ has the following relations:

$$\mu^S(\theta) = \theta V. \tag{16}$$

Therefore, a language measure is derived from the value function.

4. Learning algorithm

In this section, we propose a learning method of the optimal supervisor S for the timed DES G. The proposed method is based on the Q-learning[12]. The Bellman optimal equation correspond to Eq.(13) is described as follows:

$$\begin{aligned} Q^*(x_i, c_i^S) &= r^*(x_i, c_i^S) \\ &\quad +(1-\theta)\sum_{j\in\hat{\sigma}_i^S}\left(P_1^*(x_i, c_i^S, \sigma_j) V^*(\delta(x_i,\sigma_j))\right). \end{aligned} \tag{17}$$

where for the state $x_k = \delta(x_i,\sigma_j) \in X$,

$$V^*(x_k) = \max_c Q^*(x_k, c), \tag{18}$$

and $Q^*(x_i, c_i^S)$ is a discounted expected total reward when the supervisor S assigns c_i^S at state x_i and continues to assign the optimal control patterns until the controlled behavior reaches a terminal state. Note that, by the result of section 3, the optimal control pattern derived from Q-learning is also the optimal with regard to the language measure.

In the learning process, the supervisor S at state x_i selects a control pattern c_i^S. Then, an event σ occurs in the DES G, and the supervisor gets a reward r and observes the new state x_k. This process starts from the initial state x_1 and ends at a terminal state. The supervisor S controls the system based on the current state of the DES G. Therefore the proposed method is a type of state feedback control.

From Eq. (17), the Q^* is determined by r^*, P_1^*, V^*. So, we introduce learning parameters r', p' and p'_f as estimated values of r, p and p_f, respectively. These values have been initialized in advance. The supervisor updates r' as follows:

$$r'(x_i, c_i^S) \leftarrow r'(x_i, c_i^S) + \alpha[r - r'(x_i, c_i^S)], \tag{19}$$

and, for all active events σ' which are permitted to occur by the selected control pattern, the supervisor updates p' and p'_f as follows: For all $\sigma' = \sigma_l$ $(l \in \hat{\sigma}_i^S)$

$$p'(x_i,\sigma') \leftarrow \begin{cases} (1-\beta)p'(x_i,\sigma') \\ \quad \text{if } \sigma' \neq \sigma \text{ and } \sigma' \text{ is not forced,} \\ p'(x_i,\sigma') \\ \quad +\beta\left[\displaystyle\sum_{k\in f_i^S} p'_f(x_i,\sigma_k)\right. \\ \quad \left. +\displaystyle\sum_{k\in\{\hat{\sigma}_i^S - f_i^S\}} p'(x_i,\sigma_k) - p'(x_i,\sigma')\right] \\ \quad \text{if } \sigma' = \sigma \text{ and } \sigma \text{ is not forced,} \end{cases} \tag{20}$$

$$p'_f(x_i,\sigma') \leftarrow \begin{cases} (1-\beta)p'_f(x_i,\sigma') \\ \quad \text{if } \sigma' \neq \sigma \text{ and } \sigma' \text{ is forced,} \\ p'(x_i,\sigma') \\ \quad +\beta\left[\displaystyle\sum_{k\in f_i^S} p'_f(x_i,\sigma_k)\right. \\ \quad \left. +\displaystyle\sum_{k\in\{\hat{\sigma}_i^S - f_i^S\}} p'(x_i,\sigma_k) - p'_f(x_i,\sigma')\right] \\ \quad \text{if } \sigma' = \sigma \text{ and } \sigma' \text{ is forced,} \end{cases} \tag{21}$$

where both α and β are learning rates.

By using r', p', and p'_f, for all control patterns c' which contain events with updated parameters by Eqs. (19), (21) and (21), Q-values are updated as follows:

$$Q(x_i, c') \leftarrow r'(x_i, c') + (1-\theta) \sum_{j \in \hat{\sigma}_i^S} \left(P'_1(x_i, \sigma_j) V'(\delta(x_i, \sigma_j)) \right), \quad (22)$$

where $Q(x_i, c')$ is the estimated Q-values and for the state $x_k = \delta(x_i, \sigma_j) \in X$,

$$V'(x_k) = \max_c Q^*(x_k, c), \quad (23)$$

and P'_1 is calculated from the estimated values p' and p'_f as follows:

$$P'_1(x_i, \sigma_j) = \begin{cases} \dfrac{p'_f(x_i, \sigma_j)}{\sum_{k \in f_i^S} p'_f(x_i, \sigma_k) + \sum_{k \in \{\hat{\sigma}_i - d_i^S - f_i^S - \{m\}\}} p'(x_i, \sigma_k)} \\ \quad \text{if } f_i^S \neq \emptyset \text{ and } j \in \hat{\sigma}_i^f, \\ \dfrac{p'(x_i, \sigma_j)}{\sum_{k \in f_i^S} p'_f(x_i, \sigma_k) + \sum_{k \in \{\hat{\sigma}_i - d_i^S - f_i^S - \{m\}\}} p'(x_i, \sigma_k)} \\ \quad \text{if } f_i^S \neq \emptyset \text{ and } j \notin \hat{\sigma}_i^f, \\ \dfrac{p'(x_i, \sigma_j)}{\sum_{k \in \hat{\sigma}_i - d_i^S} p'(x_i, \sigma_k)} \\ \quad \text{if } f_i^S = \emptyset. \end{cases} \quad (24)$$

Finally, we show the learning method of control patterns from Q-values. Let $\tilde{f}_{ij} \in [0, 1]$ be a probability that the supervisor S forces the occurrence of forcible event σ_j at state x_i, and let $\tilde{d}_{ij} \in [0, 1]$ be a probability that the supervisor S disables the occurrence of controllable event σ_j at state x_i. The supervisor S assigns the control pattern according to $\tilde{f}_{ij}$ and $\tilde{d}_{ij}$. Let $\hat{f}_i^S$ and $\hat{d}_i^S$ be the index set of the forcing pattern and the disabling pattern which maximize the Q values at state x_i respectively. Then $\tilde{f}_{ij}$ and $\tilde{d}_{ij}$ are updated as follows:

$$\tilde{f}_{ij} \leftarrow \begin{cases} \tilde{f}_{ij} + \lambda(1 - \tilde{f}_{ij}) & \text{if } j \in \hat{f}_i^S, \\ \tilde{f}_{ij} + \lambda(0 - \tilde{f}_{ij}) & \text{if } j \notin \hat{f}_i^S, \end{cases} \quad (25)$$

$$\tilde{d}_{ij} \leftarrow \begin{cases} \tilde{d}_{ij} + \lambda(1 - \tilde{d}_{ij}) & \text{if } j \in \hat{d}_i^S, \\ \tilde{d}_{ij} + \lambda(0 - \tilde{d}_{ij}) & \text{if } j \notin \hat{d}_i^S, \end{cases} \quad (26)$$

where λ is a learning rate of control patterns. The above updates increase the probability of forcing or disabling events included in the estimated optimal control pattern. Therefore, the probability of selection of the pattern increases.

5. Conclusion

In this paper, we considered a supervisory control problem of the timed DESs by Brandin and Wonham, and proposed the optimal supervisory control method based on reinforcement learning. The supervisor learns the optimal control pattern with regard to the language measure.

Improvement of the algorithm for the large scale problem and extension to hybrid systems are future works.

References

[1] C. G. Cassandras and S. Lafortune, Introduction to Discrete Event Systems, 2nd ed, Kluwer Academic Pub., 2007.

[2] P. J. Ramadge and W. M. Wonham, "Supervisory control of a class of discrete-event processes" SIAM Journal on Control and Optimization, Vol. 25, No. 1, pp. 206–230, 1987.

[3] B.A. Brandin and W.M. Wonham, "Supervisory Control of Timed Discrete -Event Systems,"IEEE Trans. Automatic Control, Vol.39, No.2, pp.329–342, 1994.

[4] X. Wang and A. Ray, "Signed real measure of regular languages," Proc. of 2002 American Control Conference, pp. 3937–3942, 2002.

[5] A. Ray, V. V. Phoha, and S. Phoha, Quantitative Measure for Discrete Event Supervisory Control, Springer, 2005.

[6] T. Yamasaki and T. Ushio, "Supervisory Control of Partially Observed Discrete Event Systems based on a Reinforcement Learning," Proc. of SMC'03, pp.2956–2961, 2003.

[7] T. Yamsaki and T. Ushio, "Decentralized Supervisory Control of Discrete Event Systems Based on Reinforcement Learning," IEICE Trans. Fundamentals, Vol. E88-A, No. 11, pp. 3045-3050, 2005.

[8] T. Yamasaki, K. Taniguchi and T. Ushio, "Reinforcement Learning of Optimal Supervisor Based on Language Measure," Proc. of CDC'05, pp.126–131, 2005.

[9] Y. Murata and T. Ushio,"Optimal Scheduling of Periodic Tasks in Soft Real-Time Systems Using Language," SICE-ICASE2006, pp.1110-114, 2006.

[10] I. Chattopadhyay and A. Ray, "Renormalized Measure of Regular Languages," International Journal of Control, Vol. 79, No. 9, pp. 1107-1117, 2006.

[11] I. Chattopadhyay and A. Ray, "Language-Measure-Theoretic Optimal Control of Probabilistic Finite-State Systems," International Journal of Control, Vol. 80, No. 8, pp. 1271-1290, 2007.

[12] R. S. Suton and A. G. Barto, Reinforcement Learning, MIT Press, 1998.