ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 10, No. 2, 2015, *pp*. 106-113

Robust Reinforcement Learning Decoupling Control Based on Integral Quadratic Constraints

Wang Teng *
School of Computer, South China Normal University, Guangzhou 510631, China
(*Received August 07, 2013, accepted January 06, 2015)*

Abstract. In order to keep stable in reinforcement learning process, a novel robust reinforcement learning decoupling control (RRLDC) based on integral quadratic constraints（IQC）is presented in this paper. It composes of a linear model to approximate the nonlinear plant, a state feedback K controller to generate the basic control law, and an adaptive critic unit to evaluate decoupling performance, which tunes an actor unit to compensate decoupling action and model uncertainty as well as system nonlinearity. By replacing nonlinear and time-varying aspects of a neural network and model uncertainty with IQC, the stability of the control loop is analyzed. As a result, the range of the adjusted parameters is found within which the stability is guaranteed, the control system performance is improved through learning and the algorithm convergence speed is accelerated. The proposed RRLDC is applied to gas collector pressure control of coke ovens. The simulation results show the proposed control strategy can not only obtain the good performance, but also avoid unstable behavior in learning process. It is an effective multivariable decoupling control method for a class of strong coupling systems such as the gas collector pressure control of coke ovens.the effectiveness of proposed control strategy for the collector gas pressure of coke ovens

Keywords: reinforcement learning, decoupling control, integral

1. Introduction

Reinforcement learning (RL) is a kind of machine learning method which can be applied to solve optimal control problems[1].

Although classical reinforcement learning algorithm does not need environment models such as Sutten's TD (temporal difference) algorithm, Watkin's Q-learning algorithm and AHC (adaptive heuristic critic), there exist problems of slow convergence speed and low convergence precision. Environment models added can increase convergence speed. The nature of reinforcement learning is exploration and exploitation. Successful applications often allow systems to accumulate experience, learn from failures, and eventually succeed. However, it is difficult to ensure system stability and tracking performance in reinforcement learning process, which may greatly limit its applications in complex industrial process control.

Robust control theory is applied to uncertain systems to analyze their stability and design a controller with certain performance[2]. It has a unified framework including gain margin, phase margin, tracking, noise disturbance poise and calm concept. Most control system design methods based on robust control theory provide strong stability guarantees and certain performances. Hence, Reinforcement learning combining with robust control theory has become a hotspot in the past decades. Kretchmar et al combined robust control theory with reinforcement learning[3]. Morimoto and Doya[4] applied robust control theory to improve the performance of a reinforcement learning system under certain disturbances. Perkins and Barto[5] used Lyapunov principles to design stable controllers and achieved a level of desired performance. Jose[6] presented a series of reinforcement learning algorithms which could learn quickly, generalize properly over continuous state spaces and be robust to a high degree of environmental noise. Dong[7] proposed a novel quantum-inspired reinforcement learning algorithm for navigation control of autonomous mobile robots. This approach was then applied to navigation control of a real mobile robot. The simulation and experimental results show its effectiveness.

In the coking process, the stability of gas collector pressure has influence on coke ovens' quality, life-time and their production environment. Also, its control has a direct impact on the operating conditions of the whole coke oven system. However, the control system of gas collector pressure is a complicated multivariable, nonlinear, time-varying and big time-delay control object, which has many disturbance factors and the

* Corresponding author. Tel.: +86-15360060605.
E-mail address of corresponding author Wang Teng: towangteng@263.net

coupling phenomenon is serious. Therefore, it is difficult to establish a mathematical model to reflect collector gas pressure accurately and obtain desired control effect using traditional control methods[8].

In this paper, a novel intelligent decoupling control method based on IQC reinforcement learning is proposed and an actor-critic architecture is adopted in the RRLC, which uses a neural network to generate the decoupling control compensation and a reinforcement learning automata (RLA) to learn the neural network parameters. Meanwhile, reinforcement learning to find the optimal solution is integrated with domain knowledge available to analyze the system stability. The strategy is used for decoupling control of coke ovens plant. The simulation results show its effectiveness for gas pressure control of coke ovens.

2. The methods

A reinforcement learning algorithm

The essence of reinforcement learning is how an agent takes actions in environment so as to maximize long-term reward. The Agent consists in critic unit and actor unit. In the process of information interaction, the Agent receives environment state s and reward signal r. At the same time, the critic unit evaluates the control performance and guides the actor unit to find a policy that maps state s to action a. This action output will lead to environment change to create a new state. Every time, the choice behavior principle for the Agent is to maximize long-term reward, determine whether a new state meets the learning goals and generate a corresponding TD error signal to adjust actor unit strategies. After repeated trial and error, the Agent can finally obtain optimal behavioral strategies and complete the learning task.

B RL-based Decoupling control system architecture

RL-based decoupling control system is an actor-critic architecture, which has five components: adaptive critic evaluation element (ACE), associative search element (ASE), system model (SM), state feedback controller K, IQC stability analysis and optimal element. The system model approximates the linear part of plant. The state feedback K controller is used in the control loop giving the system a guaranteed initial performance. The ACE learns to evaluate the system performance and tune the ASE to learn nonlinear decoupling compensation mapping. The input of actual control is generated by summing the output of controller K and ASE.

Through replacing nonlinear and time-varying aspects of a neural network and model uncertainty with IQC, the stability is guaranteed in robust reinforcement learning process even as the neural network is being trained.

C. STATE FEEDBACK K

In fact, the coke oven system is a nonlinear and time-varying system. It is sure that there exist error and residuals in statistics, if a linear system, as shown in Eq(1), is used to model it.

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned} \tag{1}$$

Where y is controlled system output, X is state of system, A, B, C, D is matrix of system.

The quadratic constraint J, as performance index, is introduced to evaluate the LTI system.

$$J = \int_0^\infty [x^T(t)Q(t)x(t) + U^T(t)R(t)U(t)]dt \tag{2}$$

H function is got when minimize J [2]:

$$\frac{\partial H}{\partial u} = 0 \Rightarrow u^*(t) = -R^{-1}B^T Px(t) \tag{3}$$

Where P is got from Riccati matrix equations. Suppose $K = R^{-1}B^T P$, then the basic control law $-Kx(t)$ is obtained.The overall output, as mentioned above, is given by

$$u(t) = -Kx(t) + \tau(t) \tag{4}$$

Where decoupling compensation τ is the ASE output adding random exploration.

D. ACE AND ASE

Here, ACE and ASE are based on neural network which are standard three-layer feed-forward network with hyperbolic tangent nonlinear activation function for hidden layer and linear output layer. According to the state of the environment and instantaneous signals, ACE learns to predict the long-term reward values, through which a set of action is evaluated for the control object. ASE produces control action. Meanwhile, a reinforcement signal is obtained and used to update the weights of ACE and ASE neural network. When a state is changed into another state by taking actions, the output of ACE can be used to evaluate the previous actions. If the reward is better than ACE prediction, it means the Agent performance is improved and the probability of choosing this action can be increased, and vice versa.

The instantaneous error signal can be obtained as follows:

$$E(t) = x^T Qx + u^T Ru \tag{5}$$

Where *E(t)* is the performance signal. In this paper, the reward function is defined by external reinforcement signal *r*:

$$r(t) = -|E(t)|. \tag{6}$$

Let V be the predicting reinforcement signal by ACE, then the form of reinforcement learning signal is given by:

$$\hat{r}(t) = \begin{cases} 0 & when\ start \\ r(t) - v(t-1) & faiture\ state \\ r(t) + \lambda v(t) - v(t-1) & otherwise \end{cases} \tag{7}$$

Where λ is discount factor and 0 ≤λ≤ 1. In this paper $\lambda=0.9$. $\hat{r}(t)$ is temporal difference (TD) error. Taking the predicting reinforcement signal as an additional learning signal leads to faster and reliable learning control. ACE determined by K and ASE evaluates the "goodness" of the control actions and fine tune the weight vectors. The formulae for updating the weights of ACE and ASE are shown in section IV.

3. IQC stability analysis

Integral quadratic constraints (IQC)[6],[7] are a tool that is used to verify the stability of the uncertain system. As shown in Fig 1, the block M is a LTI system, the block △ is an uncertain block, and the feedback interconnection is considered.

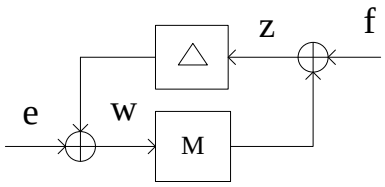


Figure 1 feedback system

IQC is an inequality describing the relationship between two signals *w* and *z*, which are characterized by matrix Ⅱ

$$\int_{-\infty}^{+\infty} \begin{vmatrix} Z(j\omega) \\ W(j\omega) \end{vmatrix}^* \prod(j\omega) \begin{vmatrix} Z(j\omega) \\ W(j\omega) \end{vmatrix} d\omega \geq 0 \tag{8}$$

Z and *W* are the Fourier Transforms of *z(t)* and *w(t)* respectively.

The system shown in figure 3 can be described as follows

$$\begin{aligned} z &= Mw + f \\ w &= \Delta(z) + e \end{aligned} \tag{9}$$

Also, the basic IQC stability theorem can be stated as follows:

Assume that:

The interconnection of M and △ is well-posed.

The IQC defined by Ⅱ is satisfied.

There exists ε >0 that for all ω , the fomula (11) is derived:

$$\begin{bmatrix} M(j\omega) \\ I \end{bmatrix}^{*} \prod(j\omega) \begin{bmatrix} M(j\omega) \\ I \end{bmatrix} \tag{10}$$

Then the feedback interconnection of M and △ is stable.

Firstly, Δ is described by IQCs as accurately as possible, and the class $\prod_{\Delta}$ of all rational Hermitian matrix functions that define a valid IQC for a given Δ is convex. When Δ composes of several simple blocks, the corresponding IQCs can be generated by convex constraints combining with these simple blocks, and the search for a suitable $\prod$ restricted to a finite-dimensional subset of $\prod_{\Delta}$ can be carried out by numerical optimization,. Then, $\prod$ can be written on the form

$$\prod(j\omega) = \sum_{q=1}^{q=q_0} x_q \prod{}_q(j\omega) \tag{11}$$

Where x_q are positive real parameters. There exist Matrix A(nxn), B (nxm), symmetric real matrices F1,…Fq0(n+m)x(n+m) that the following formula can be obtained

$$\begin{bmatrix} M(j\omega) \\ I \end{bmatrix}^{*} \prod(j\omega) \begin{bmatrix} M(j\omega) \\ I \end{bmatrix} \tag{12}$$

$$= \begin{bmatrix} (j\omega I - A)^{-1} B \\ I \end{bmatrix}^{*} F_q \begin{bmatrix} (j\omega I - A)^{-1} B \\ I \end{bmatrix} \tag{13}$$

By application of the Kalman–Yakubovich–Popov lemma, the formula (13) is equivalent to formula (14), if a symmetric matrix $P=P^T$ is existed

$$\begin{bmatrix} PA + A^T P & PB \\ B^T P & 0 \end{bmatrix} + \sum_{q=1}^{q=q_0} x_q F_q < 0 \tag{14}$$

Hence the search for $\prod$ that satisfying formula (14) can take the form of a convex optimization problem defined by a linear matrix inequality (LMI) with the variables x_q, P . This can be solved efficiently by the recently developed numerical algorithms [9], [10].The more detail and procedure are discussed in Ref[4][9].

A library of IQCs for common uncertainties is available in Matlab[9] and more complex IQCs can be constructed using the basic IQCs. To analyze the stability of the feedback loop, an IQC is placed to represent the range of parameters in actual system.

The non-LTI function φ is used to cover a time varying weight updated in ASE, which accounts for the change in the weight as the network learns. The neural network learning rate is used to determine the bounding constant α, and the algorithm looks for the largest allowable β to make sure the system's stability[4].

At first, a search process is performed for current neural network weight values using IQC to find the boundaries within which the system is stable. This defines a known "stable region" of weight values. Then, the neural network is trained in these boundaries and the weights change rate is lower than a constant. The previous step is repeated and the training is continued until the termination condition is satisfied when any weight approaches the boundaries of the stability region.

4. Application to gas collector pressure control of coke covens

A. Process description and control requirements

The structure of the gas collector pressure system for coke ovens is shown in Fig. 2. The gas generated from each chambers of coke ovens is cooled to 80-90°C through the cycle of ammonia, and then sent to primary coolers through the butterfly valves, where it is further cooled to 35-40°C. Finally it will be sent to the next process by the blast blower.

The control goal is to stabilize the collector gas pressure between 80 and 120Pa. The key problem is that there exists strong coupling among two gas collectors, the coke ovens and the blast blowers, which is often adjusted by butterfly valves on the collectors, and before the blast blowers. Here, only two butterfly valves on the collectors are adjusted by the RL and the butterfly valves before blast blowers are used to keep pressure before primary coolers smooth and steady.

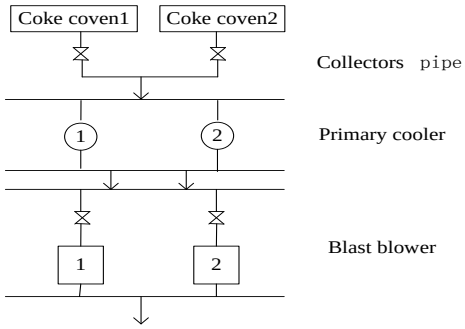


Figure.2 Scheme of collectors' pressure system

B. Reinforcement learning implementation

The coupling process is serious accompanying with uncertainty and complexity. It is difficult to establish accurate mathematics models. The traditional PID control can't deal with this nonlinear systems and often lead to serious oscillation[8]. Moreover, its control accuracy is poor, and its control effect is hard to be satisfied. So the proposed method is adapted to solve the decoupling control problem of collectors' pressure system. Its implementation process is shown in Fig. 3.

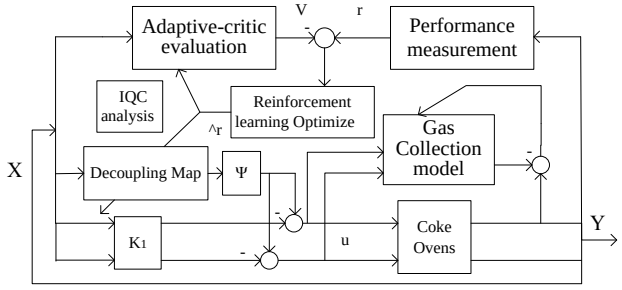


Figure.3 Reinforcement learning implementation structure

1) *Model of the plant*

At first the plant model is set up using the method described in literature[12], which consists of nonlinear differential equations. After linearization, the model can be written as a linear model adding a nonlinear and uncertain part. The model parameters are shown as follows:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} d_1 & 0 & -0.1231 & 0 \\ 0 & d_2 & 0 & -0.1231 \\ 1 & 0 & -0.3231 & 0 \\ 0 & 1 & 0 & -0.3231 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0.0215 & 0.0185 \\ 0.0215 & 0.0246 \\ 0.3231 & 0 \\ 0 & 0.3692 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0.02(1-\sqrt{x_4+1}) \\ 0.01(1-\sqrt{x_3+1}) \end{bmatrix} \tag{15}$$

Where $y_1(t) = x_3(t), y_2(t) = x_4(t)$, $u_i(t)$ donates the action output, $y_i(t)$ is pressure output. d_i donates disturbance and uncertainty.

2) *External reinforcement signal*

In this system, two external pulse disturbance at time 0 and time *tl* are acted on the gas collectors, then the performance index is considered as external reinforcement signal, quadratic index *Ei* is used to measure control performance:

$$r(t) = -\left| e^T Q e + u^T R u \right| \tag{16}$$

where $e = [e_1 e_2]$, $e_i(t) = y_i - y_i^d$, *i=1,2*, y_i are the controlled variables, in our work it is the collectors' gas pressure, y^d are the expected value. Q and R are the weight coefficients matrix. *Q*=diag[0.05,0.05,2,2], *R*=diag[0.05,0.05]. *d1=0.012, d2=0.015* and when t is between [3,3.2],[8,8.5], The feedback matrix is given by:

$$K=\begin{bmatrix}1.2060 & 0.9519 & 6.3017 & -0.0199\\ 0.9508 & 1.2040 & -0.0247 & 6.3030\end{bmatrix}$$

3) learning algorithm for ACE and ASE

The forward networks structure is shown in Fig. 4. The output of ACE is the evaluation signal *V(t)*, which is the estimated values of long-term reword. The output of ASE is the decoupling compensation (Represented by a1, a2). The same state vectors *X*(*t*) are used for ASE and ACE input, and all the networks use the same input layer(wi), hidden layer:

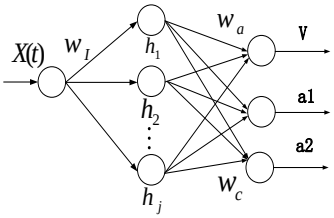


Fig. 4 networks structure

W_I are the hidden layer input weights, w_a,w_b,w_c are the output layer weights The arctan function is used in hidden nodes.

$$h_j=\arctan(w_I X) \tag{17}$$

Where j=1,..,N_h, j=1,2,...,m. N_h is the number of neurons in hidden layer, $X=[x_1,\cdots,x_4]^T$ The output of m nodes to ASN and ACN are defined as follows:

$$V_t=\sum_{i=1}^{N_h} h_j w_a \tag{18}$$

$$a_{1t}=\sum_{i=1}^{N_h} h_j w_b \tag{19}$$

$$a_{2t}=\sum_{i=1}^{N_h} h_j w_c \tag{20}$$

w_a, w_b, w_c are hidden layer to output layer weights respectively. After a_i is got, a maximum probability action is selected by Random Action Selection Unit. The actual decoupling compensation action applied to the gas collector plant can then be computed as:

$$\tau_i=\psi(a_i,\sigma)+u_{Ki} \tag{21}$$

Where Ψ is a normal distribution function with mean a_i, which provides the required exploration around a_i. And the variance σ is given by:

$$\sigma=\frac{k_1}{1+\exp(k_2 V)} \tag{22}$$

Where k_1 is 0.05 and k_2 is 5. If the predicted reinforcement V is high, the network performance is well, and hence the amount of exploration should be minimized. If the V is low, the network performance is poor and the more exploration is needed to find better actions.

In the critic network, supervised learning is used. The objective function is the differentiation function between the prediction and the actual long-term reward which is defined as (7) .

ACE weights w_c are updated as follows:

$$w_{Ii}(t+1)=w_{Ii}(t+1)-\delta\hat{r}(t)x_i(t) \tag{23}$$

$$w_{ai}(t+1)=w_{ai}(t+1)-\delta\hat{r}(t)h_i(t) \tag{24}$$

Where δ is the learning factor.

The action recommended by (21) can then be written as

$$\frac{\partial In\psi}{\partial \tau} = (\frac{a - \tau}{\sigma^2}) \tag{25}$$

ASE weights w_a are updated as follows:

$$w_{ai}(t+1) = w_{ai}(t) - \eta(\frac{a_1 - \tau_1}{\sigma^2})h_i(t) \tag{26}$$

$$w_{bi}(t+1) = w_{bi}(t) - \eta(\frac{a_2 - \tau_2}{\sigma^2})h_i(t) \tag{27}$$

η is the learning rate，

4） stability analysis with IQC

For stability analysis, the nonlinear function is replaced by an IQC. The nonlinearity meets the conditions of the odd slope-restricted nonlinearity IQC

Formula (17) can be rewritten as:

$$h_j = \frac{\arctan(w_I X)}{w_I X} w_I X = \Phi(w_I X) w_I X \tag{28}$$

As described in section 3, the arctan function, namely the active function of hidden nodes in neural network, meets the conditions of the odd slope-restricted nonlinear IQC:

$$\arctan(-x) = -\arctan(x) \qquad 0 \le (\arctan(x_1) - \arctan(x_2))(x_1 - x_2) \le (x_1 - x_2)^2 \tag{29}$$

So the matrix Φ can be replaced by an uncertainty function and the bounded odd slope nonlinearity IQCs can be used to describe it. The matrix Φ can be replaced by diagonal matrix of these bounded odd slope nonlinearity IQCs with an appropriately dimensions. In this paper if the weight vectors are satisfied:

$$|w_c| \le \beta \tag{29}$$

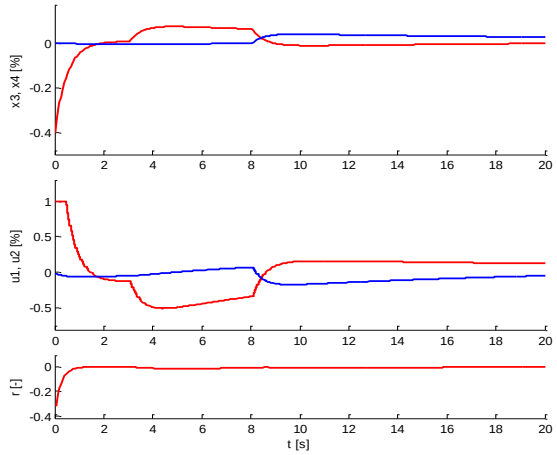
$$\left|\dot{w}_c\right| \le \zeta \tag{30}$$

This linear gain block with slowly time-varying can be replaced by slowly time-varying real scalar IQCs.

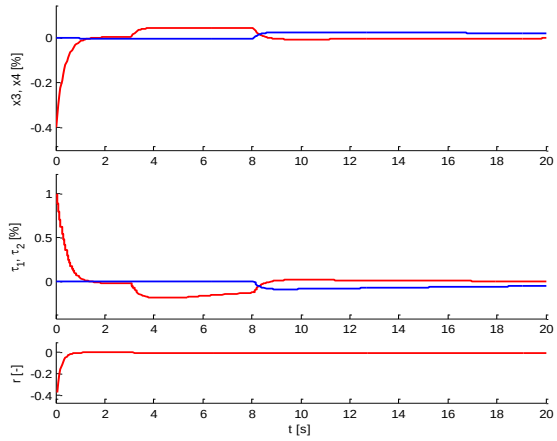
Others uncertainties such as model uncertainty are also replaced by appropriate IQCs and carried out a search procedure in fig 4. This procedure includes using IQC analysis to find bounds of weight values within which the system is stable.

C. The simulation results

The simulation results are shown in Fig. 5. There exists disturbance 0.15 and 0.1 to the *x3* and *x4* respectively. The top diagram in Figure 5 shows the system corresponding to the state controller K and neural controller before learning . The bottom diagram shows the same system after learning. It can be seen that the response is quickly and the coupling between the two gas pressures are weakened greatly.



(a)



(b)

Figure 5 Simulation results of RRLC (a) before learning, (b) after learning

5. Conclusion

It is very difficult to guarantee the control stability while RL is learning. This paper presents a new control strategy for gas collector pressure control of coke ovens based on robust reinforcement learning, which adopts an actor-critic architecture and uses IQC analysis to implement the robust reinforcement learning. The use of IQC enables the system to cope with difficulties in the stability analysis with uncertainty. The robust reinforcement learning can accelerate the convergence speed on actor-critic learning. It is very suitable for multivariable decoupling control. Further work will be involved in distributed reinforcement learning to reduce cycle time and improve the system performance using multi-agent cooperation and collaboration.

ACKNOWLEDGMENTS

Project supported by the National Natural Science Foundation of China (61074067, 21106036); Guangdong Provincial Natural Science Foundation (2014A030310153);Key technology R&D program of Hunan province(2014FJ2018).

6. References

[1] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press, 1998

[2] Kemin Zhou and John C. Doyle. Essentials of Robust Control. Prentice Hall, 1998.pp:125-156

[3] R. Matthew Kretchmar, Peter M. Young, Charles W. Anderson, et al, Robust Reinforcement Learning Control with Static and Dynamic Stability. Delnero Colorado State University. Technical Report CS-00-102.2001,5(30). pp:8-11

[4] J. Morimoto and K. Doya, Robust reinforcement learning, Neural Comput.,vol.17,pp.335–359,2005.

[5] Stephan V, Debes K., Gross H-M. A new control scheme for combustion processes using reinforcement learning based on neural networks, International Journal of Computational Intelligence and Applications 2001, 1(2).pp:121-136

[6] Jose Antonio Martin H, Javier de Lope, Dario Maravall et al. Robust high performance reinforcement learning through weighted k-nearest neighbors[J]. Neurocomputing, 2011,74(8):1251-1259

[7] Dong, D.,Chen, C.,Chu, J. et al.Robust Quantum-Inspired Reinforcement Learning for Robot Navigation [J]. IEEE/ASME transactions on mechatronics, 2012, 17(1): 86-97.

[8] Wang, Xin, Yang, Chunhua, Qin, Bin. "Decoupling Control Using a PSO-Based Reinforcement Learning", 2009 International Conference on Computational Intelligence and Natural Computing，2009

[9] Alexandre Megretski and Anders Rantzer. System analysis via integral quadratic constraints. *IEEE Transactions on Automatic Control*, 42(6):819–830, June 1997

[10] Alexandre Megretski, Chung-Yao KAO, Ulf Jonsson, and Anders Rantzer. A Guide to IQCβ: Software for Robustness Analysis. MIT/ Lund Institute of Technology, http://www.mit.edu/people/ameg/ home.html, 1999.

[11] C W Anderson, P M Young, M R Buehner et al. Robust reinforcement learning control using integral quadratic constraints for recurrent neural networks. IEEE Trans Neural Netw. 2007 Jul;18(4):993-1002

[12] Bin Qin; MAS-Based Intelligent Pressure Decoupling and Coordinate Control of Gas Collectors in Coke Ovens, Ph.d Dissertation. Central South University,2006. (in Chinese)