Robot reinforcement learning accuracy-based learning classifier systems with Fuzzy Policy Gradient descent(XCS-FPGRL)

Jie SHAO, jingru YU

Zhengzhou Chenggong University of Finance and Economics,Department of information Engineering 451200 zhengzhou China

Keywords: Convergence, Rrobot, Reinforcement learning, Accuracy-based learning classifier system with Gradient descent (XCS-FPGRL)，XCS（Accuracy-based learning classifier system）

Abstract. This paper presented a novel approach XCS-FPGRL to research on robot reinforcement learning. XCS-FPGRL combines covering operator and genetic algorithm. The systems is responsible for adjusting precision and reducing search space according to some reward obtained from the environment, acts as an innovation discovery component which is responsible for discovering new better reinforcement learning rules. The experiment and simulation showed that robot reinforcement learning can achieved convergence very quickly.

1 Introduction

Robot technology has already been widely used in many fields[1-3]. Urged by investigation and application, research of robots reinforcement learning has been a hot field. In some task orientation applications, robot reinforcement learning convergence and efficient navigation are key technologies of robot navigation, and also the requirement of the rapid development of robot system navigation technology [4-9].

Since most reinforcement learning algorithms are based on the assumption of completely unknown environment model, through the study of the system and the environment interaction to achieve sequential optimization decision.

To solve the complex problems often learning cost, slow speed of convergence, and many complex real-world applications are often not model is completely unknown, at least there are many can learn from the prior domain knowledge.

On the one hand can enhance learning computational search algorithm quickly focused on optimization strategy and value function space; on the other hand, can also make use of prior knowledge in the calculation of the search to accelerate the convergence process

Therefore, the fusion of prior domain knowledge to enhance learning algorithm and theory has become the important development trend of enhanced learning research.

One of the research hotspots is a hybrid algorithm combining learning and supervised learning enhancement research, namely the use of supervised learning results to constrain the reinforcement learning problem space, the reinforcement learning algorithm can lower cost under the conditions of learning space focused on spatial behavior strategy better, showed learning in reinforcement learning the possibility and advantages of.

2 XCS-FPGRL

The XCS-FPGRL*(Accuracy-based learning classifier system with Gradient descent)* classifier system is a learning classifier systems (LCS) that evolves its classifier by an accuracy-based fitness approach [10-14].

One of the research hotspots is a hybrid algorithm combining learning and supervised learning enhancement research, namely the use of supervised learning results to constrain the reinforcement learning problem space, the reinforcement learning algorithm can lower cost under the conditions of learning space focused on spatial behavior strategy better, showed learning in reinforcement learning the possibility of combining with discrete behavior fuzzy policy gradient reinforcement learning (fuzzy policy gradient reinforcement learning, FPGRL) reinforcement learning to adjust

the parameters of the fuzzy rules by policy gradient. Because of the fuzzy rules is difficult to develop, so the use of reinforcement learning methods, through interaction with the environment, to adjust the parameters of fuzzy rules, fuzzy inference system as to the policy gradient reinforcement learning function approximators, so as to realize the action selection, selection of probability. Advantage

The state set $s=(s_1,s_2,\cdots,s_k)$, each rule incentive intensity:

$$\alpha_i(s)=\prod_{j=1}^{k}u^i{}_j(s_j) \tag{1}$$

The j dimension of output of the rule base:

$$Q_j=\sum_{i=1}^{N}(a_i(s)\times q^i{}_j) \tag{2}$$

Tthe j behavior of the probability of selection:

$$F_j=\frac{\exp(Q_j(s))}{\sum_{i=1}^{m}\exp(Q_i(s))} \tag{3}$$

3 Robot reinforcement learning

3.1 Improved crossover operator based on XCS

Each XCS classifier contains the usual condition, action, and reward prediction parts. Swarm robots reinforcement learning convergence is inseparable from environmental information and it is particularly important that how to obtain timely accurate environment information through reinforcement learning. Therefore, gradient descent method mapped to learning classifier system and integrated support vector machine algorithm, a new algorithm of learning classifier system based on gradient strategy in application of robots reinforcement learning is proposed.

Step1：The policy parameter vector in XCS , the classifier is represented as <conditions, policy parameters> .

Step2 ： Select $p_1=(r_1,s_1)$ and $p_2=(r_2,s_2)$, Conditions and policy parameters is cross-evolution, $r_i=(a_1^{p_i},a_2^{p_i},\cdots a_l^{p_i})$ and $s_i=(\delta_1^{p_i},\delta_2^{p_i},\cdots\delta_l^{p_i})$ is produced.

Step3：each new classifier o_i produced a new strategy parameter vector $s_i=(\delta_1^{o_i},\delta_2^{o_i},\cdots,\delta_l^{o_i})$, $\delta_j^{o_i}$ is random parameters , $\delta_j^{o_i}\in[c^i_{\min}-I_i\alpha,c^i_{\max}+I_i\alpha]$, and $c^i_{\min}=\min(\delta_i^{p1},\delta_i^{p2})$, $c^i_{\max}=\max(\delta_i^{p1},\delta_i^{p2})$, $I_i=(c^i_{\min},c^i_{\max})$.

3.2 Fitness function design

We think it as the main elements of reinforcement learning that robot can avoid obstacles in dynamic or static. So We design the fitness function by whether warm robots is within safety limits and the robot strength .

3.2.1 Whether within the security

We considered all obstacles as particles, each obstacle has a safe radius, if the distance between robot and obstacle is greater than the safety radius, then considered to be safe; if it is less than the

safety radius, then that is not safe. Relationship between the radius R and distance d as follows:

$$fit1 = \begin{cases} 0 & \mathrm{d} \geq \mathrm{R} \\ -1 & \mathrm{d} \leq R \end{cases}$$

(4)

where $d = \sqrt{(x_o - x_r)^2 + (y_o - y_r)^2}$, Obstacle coordinates（$x_o$,$y_o$）, robot current coordinates（ x_r , y_r）. $fit1$ value indicates the robot particle and obstacles are in a safe distance , then its fitness is 0, if the robot particle and obstacles are not in a safe distance, then the fitness is -1.

3.2.2 The fitness function for robot strength

$$fit2 = S_i(t+1)$$

(5)

3.2.3 Probability success rating fitness function of the robot system

The multi-robot system does not reach the target position before, individual robot to move within the region, at the moment, the robot to the target position prediction own pre-close to the desired speed and azimuth of the target is calculated, and the results posted via the wireless network and to otherrobot upgrade for leadership robot. Other robots calculated according to the received information on its own position and movement trend, and also calculate the target and other robots and their relative position, the case according to the other robots, the speed and azimuth of the own pre close to the ideal target for finallevel programs and leading robot using weighted probability of success the first establish the position estimation equation as follows:

$$\begin{cases} \left|(x_i + v_i t\cos\theta_i) - (x_g + v_g t\cos\theta_g)\right| = \varepsilon_{i,x} \\ \left|(y_i + v_i t\sin\theta_i) - (y_g + v_g t\sin\theta_g)\right| = \varepsilon_{i,y} \\ \varepsilon^2{}_{i,x} + \varepsilon^2{}_{i,y} = \overline{\sum{}^{'2}{}_0} \end{cases}$$

(6)

Where x_i, y_i, v_i, θ_i is respectively the robot position ,expected speed and azimuth; x_g, y_g, v_g, θ_g is respectively the robot the target position, speed and direction; Σ′0 is The target robot mobile security Radius; $\varepsilon_{i,x}$ and $\varepsilon_{i,y}$ is the deviation of the two-dimensional component.

Leadership robot x_{leader}, y_{leader}, v_{leader}, θ_{leader} into the above, and obtained t_{leader} on behalf into the following formula This system feasibility probability as follows:

$$p_{system} = p(\varepsilon^2{}_{i,x} + \varepsilon^2{}_{i,y}) \leq \overline{\sum{}^{'2}{}_0} \mid t_i \leq t_{leader}$$

(7)

$$fit3 = \begin{cases} 1, P_{system} \geq 0.65 \\ 0, P_{system} < 0.65 \end{cases}$$

(8)

where $P_{system} \geq 0.65$, indicating the overall planning of the swarm robots system is feasible; $P_{system} < 0.65$, indicating the overall planning of the swarm robots system is not feasible.Taking these factors, the integration fitness function of swarm robots reinforcement learning as follows:

$$f = (1 + fit1) * fit2 * fit3 * F_j$$

(9)

3.3 Convergence strategy of swarm robots path planning

The XCS classifier system is an LCS that evolves its classifier by an accuracy-based fitness approach. Each XCS classifier contains the usual condition, action, and reward prediction parts. Complementary, XCS-FPGRL contains a prediction error estimate and a fitness estimate, which represents the relative accuracy of a classifier.

XCS can find a set of learning rules through interaction with the environment. These rules can be used to guide the robot collision avoidance, and can provide real-time, dynamic feedback for the robot. The warm robots can autonomous learn optimal convergence strategy.

Step1:The optimal learning strategies as the XCS initial rule set, and randomly perform one of the learning strategies.

Step2:Condition part of the message in the message list is compared with the current regulations, matched regulations is put into the matching rule set.

Step3:Selecting the appropriate classifier from the matching rules, and sent these rules to the effectors to guide the robot to generate the corresponding planning action, and to determine the convergence effect based on the feedback value.

Step4: Repeat Step2-Step 3 until the matching rule set is empty or rule discovery mechanism is triggered.

Step5:The rule discovery system using genetic algorithms and rules covering algorithm to construct the new rules.

Step6:Merge the new rules, so that it can be summed up the previous two samples. The new learning of individual rules are sent the public rule set for the other robot XCS-FPGRL to generate new rules to share.

Step7:If robot does not meet the convergence effect, return to Step2.

4 Experiments and Simulation

Suppose the activities region of swarm robots is a rectangular area, O1-O3 are three dynamic obstacles, the rest is static obstacles in the region. Fig.1 shows the swarm robots reinforcement learning based on XCS in a static environment. Fig.2 shows Multi-robot trajectory in the U-shaped environment based on XCS-FPGRL, effectively improving the learning convergence speed.

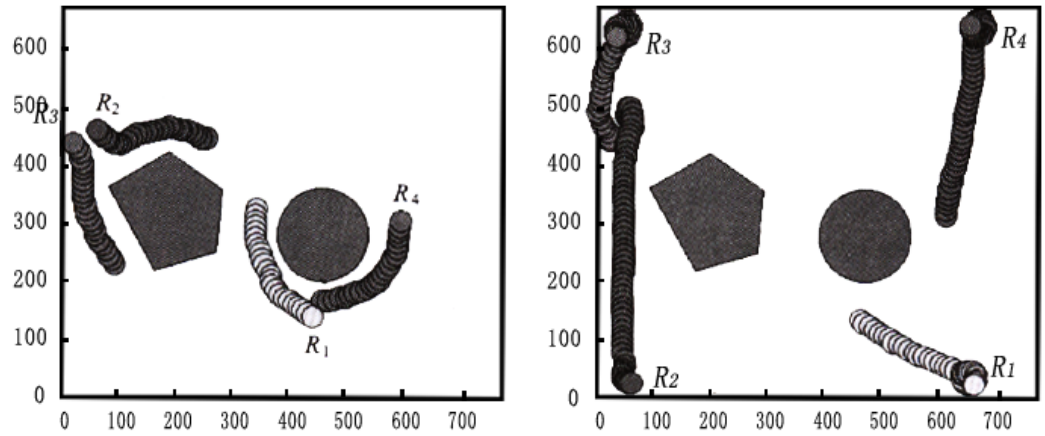

Fig. 1 Robot reinforcement learning in a static environment based on XCS

Under normal circumstances, the performance of a reinforcement learning algorithm needs two aspects to determine, one is convergence of algorithm, the other is convergence speed. Figure 2 shows Multi-robots in the narrow context of the U-shaped trajectory. Dynamic obstacles (No. 1-3) were in their narrow U-shaped environment for up and down movement, four were successfully reach the ultimate goal of the robot point G point.

From Fig 2 and Fig3 we can see algorithm convergence trend after a certain number of learning. After a certain number of learning, the oscillation amplitude is very small, we can consider the

approximation to the convergence. Fig3 is reinforcement learning based on XCS-FPGRL, whether it is the robot individual reinforcement learning or multi-robot learning, all robot can receive the desired learning convergence curve in the short period of time.

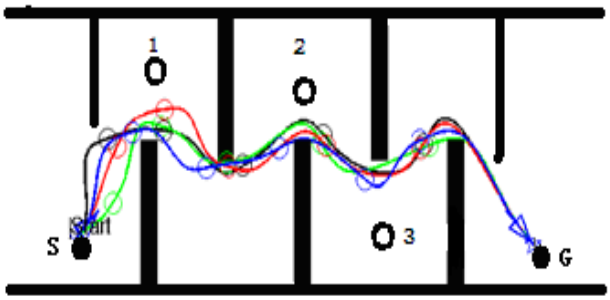

Figure2 Multi-robot trajectory in the U-shaped environment based on XCS-FPGRL

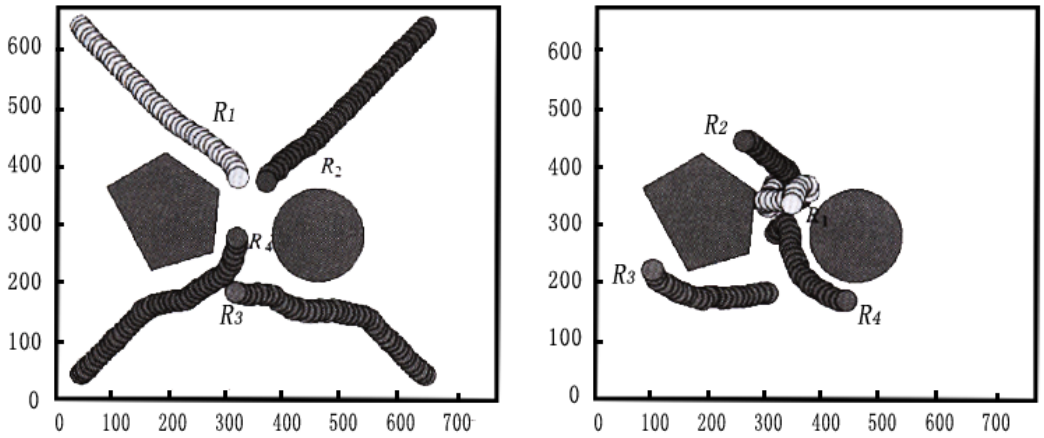

Figure 3 Multi-channel robot trajectories in narrow channel environment XCS-FPGRL

Figure 3 is the in the multi-robot trajectory in narrow channel environment. Dynamic obstacles (No. 1-3) were down as a straight line in their environment, all involved in the planning of the robots have achieved satisfactory simulation curve. Presented in this paper XCS-FPGRL based integration algorithm, because all XCS-FPGRL have strong ability to forecast returns, all the stability of the robot can achieve a satisfactory convergence effect for multi-robot trajectory in the narrow dynamic environment.

Conclusions and Future works

This paper presented a novel approach to solve the problem of swarm robot reinforcement learning convergence. XCS-FPGRL is an accuracy-based machine learning system that combines covering operator and genetic algorithm. The covering operator is responsible for adjusting precision and reducing search space according to some reward obtained from the environment. The genetic algorithm acts as an innovation discovery component which is responsible for discovering new better reinforcement learning rules. The advantages of this approach is its accuracy-based representation, that can be easily reduce learning space, improve online learning ability, robustness due to the use of genetic algorithm. Simulation indicated that the accuracy-based learning classifier system used in the swarm robot's reinforcement learning convergence is effective, and swarm robots can achieve stably convergence very quickly.

Acknowledgement

The authors would like to thank the anonymous reviewers and the editor for their helpful comments and suggestions. This work is partially supported by the Nature Science Foundation

(Project No. 201112400450401) and 2013 Henan College "professional comprehensive reform pilot" project and 2012 Education Department of Henan Science and Technology Research Key Project (Project No. 12B520047)

References

[1] Shao Jie , Yang Jingyu , Wan Minghua , and Huang Chuanbo “Research on Cnvergence of Multi-Robot Path Planning Basedon Learning Classif ier System[J]”. Journal of Computer Research and Development, 47 (5) : 948-955 , 2010.

[2] Lan Ting,Liu shirong. “Research on Multi-Robot robot system inspired by Biological Swarm Intelligence[J]” .Robot ,2007,29(3):298-304.

[3] Shao Jie , Yang Jingyu. “Research on convergence of robot path planning based on LCS[C]”. In Proceedings of Chinese Conference on Pattern Recognition, pp.271-276, Oct. 22-24, 2009, Nanjing, China

[4] P. W. Dixon, D. W. Corne, and M. J. Oates, “Apreliminary investigation of modified XCS as a generic data mining tool,” in Advances in Learning Classifier Systems, Germany: Springer-Verlag, 2002, vol. 2321,LNAI, pp. 133–150.

[5] M. Gemeinder, and M. Gerke, “GA-based path planning for mobile robot systems employing an active search algorithm[J]”,*Applied Soft Computing*, Vol.3, pp149-158, 2003.

[6] S. M. Baneamoon, R. Abdul Salam, A. Z. Hj. Talib, “ Learning Process Enhancement for Robot Behaviors[J]”, International Journal of Intelligent Technology, Volume 2 Number 3, ISSN 1305-6417, 2007, pp. 172-177.

[7] L.bull,M.studley,A.Bagnall, I.whittley. “Learing classifier system ensembles with rule-sharing[J]”, IEEE transactions on evolutionary computation, Vol,No.4,august 2007,pp.496-502.

[8] Baneamoon S M,Salam R A.Applying steady state in genetic algorithm for robot behaviors[C]//2008 International Conference on Electronic Design.Piscataway,NJ,USA:IEEE,2008:930-934.

[9] L. Bull. “A Simple Accuracy-based Learning Classifier System”, University of the West of England, Bristol, 2003.

[10] Y. Wang, M. Huber, V. N. Papudesi, and D. J. Cook, “User-guided reinforcement learning of robot assistive tasks for an intelligent environment,” in Proc. IEEE/RJS Int. Conf. Intell. Robots Syst., 2003, vol. 1, pp. 424–429.

[11] L. Bull and T. Kovacs, “Foundations of learning classifier systems: An introduction,” in Foundations of Learning Classifier Systems. New York: Springer-Verlag, 2005, vol. 183, pp. 1–17.

[12] P. Musilek, Sa Li, and L. Wyard-Scot, “ Enhanced Learning ClassifierSystem for Robot Navigation ”, IROS 2005, IEEE/RSJ International Conference on Intelligent Robots and Systems, Alberta, Canada, 2-6Aug. 2005, pp 3390- 3395.

[13] L.bull,J.Sha’Aban,A.Tomlinson,P.Addison.“Towards distributed adaptive control for road traffic junction signals using Learing classifier systems”. In applications of Learing classifier systems,berlin germany:Springer-verlag,2004,pp.276-299.

14] S. J. Bay, “Learning Classifier Systems for Single and Multiple Mobile Robots in Unstructured Environments[J]”, Mobile Robots X. Philadelphia,PA, Nov. 1995, pp. 88-99.