Two mode Q-learning

Kui-Hong Park **Jong-Hwan Kim**

Department of Electrical Engineering and Computer Science,
Korea Advanced Institute of Science and Technology (KAIST),
Guseong-dong, Yuseong-gu, Daejeon, 305-701, Republic of Korea
{khpark, johkim}@rit.kaist.ac.kr

Abstract- In this paper, a new two mode Q-learning using both the success and failure experiences of an agent is proposed for the fast convergence, which extends Q-learning, a well-known scheme used for reinforcement learning. In the Q-learning, if the agent enters into the "fail" state, it receives a punishment from environment. By this punishment, the Q value of the action which generated the failure experience is decreased. On the other hand, the proposed two mode Q-learning is based on both the normal and failure Q values for the selection of the action in a state-action space. To determine the failure Q value using the previous failure experience of the agent, it employs a failure Q value module. To demonstrate the effectiveness of the proposed method, it is compared with the conventional Q-learning in a goalie system to perform goalkeeping in robot soccer.

1 Introduction

Reinforcement learning is learning what to do so as to find out optimal action in each situation. In the reinforcement learning scheme, there is an interaction between the agent and the environment [1, 2]. The agent learns to select the optimal action that yields the maximum reward [3, 4]. For finding out the optimal action, the agent has to go through numerous trials and errors during the learning stage.

Q-learning is a well-known scheme in reinforcement learning [5]. It is easy to implement and is not affected by the learning policy for Q value convergence [6, 7]. Hence, it has been used in many application areas [8, 9, 10, 11]. After the Q value converges to the optimal value, the agent selects only the action with the maximum Q value in a given state. For the convergence of the Q value to the optimal value, the state-action pairs should be visited by the agent many times [12, 13]. To improve the speed of the Q-learning, several modifications based on the conventional Q-learning have been presented [14, 15, 16].

In the Q-learning a failure experience of the agent is a situation when the agent enters into the "fail" state. In that case, the environment has the fail state in its state-action space. When the agent reaches the fail state, it receives a punishment, i.e. one negative reward from the environment. By receiving the negative reward, the Q value of the action reaching the fail state is decreased. Generally, the negative reward is given to the agent due to its failure experience in the Q-learning.

In this paper, a new two mode Q-learning is proposed for improving the performance of the Q-learning using the failure experience of the agent more effectively along with the success experience. It consists of an action selection module, a normal Q value module and a failure Q value module. The action selection module and the normal Q value module are the same as those of the conventional Q-learning. In the failure Q value module, the failure Q value is calculated based on the failure probability which is determined by considering the failure experience of the agent. By using both normal and failure Q values, the action is selected in a state-action space. The effectiveness of the proposed two mode Q-learning is compared to that of the conventional Q-learning in the nondeterministic environment. As a nondeterministic environment, a soccer robot system is used to evaluate its performance against the conventional Q-learning in training a goalie robot.

Section II proposes a new two mode Q-learning. In section III, the simulation results showing the comparison between the performances of the Q-learning and the two mode Q-learning are presented. Concluding remarks follow in section IV.

2 Two mode Q-learning

In Q-learning, an agent learns through its experience in the environment. The agent that has faced many failure experiences acquires some useful knowledge to be learned from it. Due to the useful knowledge from the failure experience, the agent will have more possibility of not going into the fail state and have more chances of entering into the success state in the next trial. When the environment has several fail states, the agent may reach one of these fail states. In this situation, the agent receives punishment from the environment and restarts from the initial state for reaching the success state. The failure experience of the agent is simply reflected into the Q-value of the action by the punishment.

In this paper, the performance of the proposed Q-learning called two mode Q-learning, is improved by using failure experiences of the agent more efficiently along with the success experiences. To utilize the failure experience, a new failure Q value module is proposed. Both of the Q value from the conventional Q-learning and the failure Q value from the failure Q value module are used for action selection in the proposed scheme. The failure Q value is calculated in the failure Q value module based on the failure probability, which is obtained from the failure experience.

The terminologies for the two mode Q-learning are as follows:

- State-action trace: The state, s_j of the agent is

0-7803-7804-0 /03/$17.00 © 2003 IEEE

changed by selecting one action, a_i among the several possible actions in the current state. As a result of the action in the current state, the agent will be located in the next state. The sequence of these state-action pairs is called state-action trace.

- Step: A unit distinguishing the current state-action from the next state-action.
- Trial: There are two types of trial for arriving at the final state, as follows:
  1. Success trial: It occurs when the agent starts from an initial state and arrives at the success state.
  2. Failure trial: It occurs when the agent leaves the initial state but can not reach the success state.

2.1 Architecture of the two mode Q-learning

The new algorithm uses both of the normal Q value and the failure Q value so that it is called the two mode Q-learning. In the two mode Q-learning, two kinds of Q value are considered as follows:

- Normal Q value, Q_N: Q value from the conventional Q-learning consisting of the action selection module and normal Q value module
- Failure Q value, Q_E: Q value from the failure Q value module based on the failure experience

The architecture of the two mode Q-learning consists of the action selection module, the normal Q value module and the failure Q value module as shown in Figure 1.

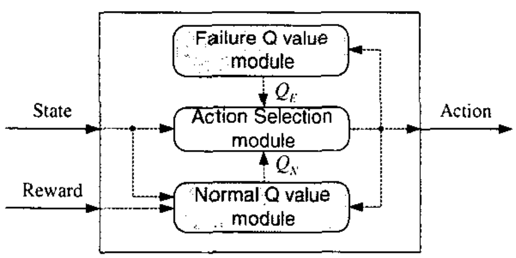


Figure 1: Architecture of the proposed two mode Q-learning

Roles of these modules are as follows:

- Failure Q value module: Calculates the failure Q value using the failure probability. The failure probability depends on the step length of the state-action trace of the failure experience, which will be described in detail in section 2.1.2.
- Normal Q value module: Calculates the normal Q value
- Action selection module: Selects an action based on the Q value, summation of Q_N and Q_E

Note that both the action selection module and the normal Q value module construct the conventional Q-learning.

The normal Q value is updated as follows:

$$Q(s_t, a_t) \longleftarrow Q(s_t, a_t)+\alpha(r_{t+1}+\gamma \max_{a'} Q(s_{t+1}, a')-Q(s_t, a_t)) \quad (1)$$

where Q is the action value, α is the learning rate, r is the reward value and γ is the discount rate.

2.1.1 Action selection

For the action selection in the Q-learning, the following Boltzmann action selection [17] is generally used for calculating the probability of action a_i being selected in state s_j:

$$p(s_j, a_i) = \frac{e^{Q(s_j, a_i)/\tau}}{\sum_{a_i \in A} e^{Q(s_j, a_i)/\tau}} \quad (2)$$

where A is an action set in the next state, τ is the temperature, which starts at high temperature at the initial stage of the learning and decreases to restrict the characteristic of exploration to a smaller zone as the number of trials increases.

In the two mode Q-learning, the action selection depends on the total value of the normal Q value ($Q_{N_{ij}}$) and the failure Q value ($Q_{E_{ij}}$). Thus, the following Boltzmann equation is employed:

$$p(s_j, a_i) = \frac{e^{Q_T(s_j, a_i)/\tau}}{\sum_{a_i \in A} e^{Q_T(s_j, a_i)/\tau}}, \qquad Q_{T_{ij}} = Q_{N_{ij}} + Q_{E_{ij}} \quad (3)$$

where $Q_{N_{ij}}$ is the Q value of action i in state j obtained by the conventional Q-learning, $Q_{E_{ij}}$ is the Q value of action i in state j obtained from the failure Q value module, and $Q_{T_{ij}}$ is the total value of these two Q values. $Q_{T_{ij}}$ decreases if the value of $Q_{E_{ij}}$ decreases, and hence the probability of action a_i being selected in state s_j is lowered.

2.1.2 Failure Q value Module

To use the failure experience more effectively in the learning process, a failure Q value module is introduced. The failure Q value module is to calculate the failure Q value by applying the failure probability to the action in the state-action trace of the previous failure trial.

The failure Q value of action a_i in a given state s_j is calculated by taking the numerator form of the Boltzmann equation, $e^{Q_{E_{ij}}/\tau} = 1 - p_{F_{ij}}$:

$$Q_{E_{ij}} = \tau\, ln(1 - p_{F_{ij}}), \quad (4)$$

where τ is the temperature value used in (3), i is the index of the action, and j is the index of the state. And $p_{F_{ij}}$ is a failure probability of the actions included in the failure trial, which is obtained from the following equation:

$$p_{F_{ij}} = f(p^N_{min_{ij}}, p^N_{max_{ij}}(k), l(c)), \quad (5)$$

where N is the index of the failure trial (the state-action trace that led the agent to the fail state), i and j are the index of the action and state, respectively, included in the failure trial N. $p^N_{min_{ij}}$ and $p^N_{max_{ij}}(k)$ are minimum and maximum probability values, respectively to be assigned by the user (see equation (7)), and k is the trial number. $l(c)$ is the step length of the state-action trace counted from the fail state. Failure probabilities are to be calculated for the actions included in the step length $l(c)$, where c is the constant value. According to the environment, the number of steps that led the agent to the fail state are varied. It means that the step length $l(c)$ of state-action trace depends on the environment. Hence, $l(c)$ can be selected as follows:

$$l(c) = \begin{cases} c\,(1 + random(0,1)) & \text{if} \quad m > 2c \\ c + 1 & \text{else if } m > c+1 \\ \min(c-1, m-1) & \text{otherwise} \end{cases} \tag{6}$$

where $random(0,1)$ is a random number generation function between 0 and 1, and m is the number of steps of the state-action trace leading to the fail state.

Figure 2 depicts an example of the state-action traces leading the agent to the success state ($(n-2)$th and $(n+1)$th trials) and the fail state ($(n-3)$th, $(n-1)$th, and nth trials) and the step lengths of the state-action traces of the failure trials that help to calculate the failure probability.

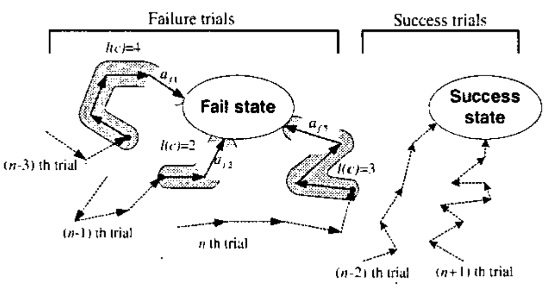


Figure 2: State-action traces of failure and success trials

In the figure, a_{f1}, a_{f2} and a_{f3} are assigned sequentially to the actions that take place just before the fail states of the $(n-3)$th, $(n-1)$th, and nth trial, respectively. This figure shows the cases of $l(c) = 4$, $l(c) = 2$, and $l(c) = 3$ in the $(n-3)$th, $(n-1)$th, and nth trial, respectively. The failure probability applied to the actions is calculated as follows:

$$\begin{aligned} p_{F_{ij}} &= f(p^N_{min_{ij}}, p^N_{max_{ij}}(k), l(c)) \\ &= p^N_{min_{ij}} + \frac{p^N_{max_{ij}}(k) - p^N_{min_{ij}}}{l(c)}\,(i - (m - l(c))), \\ &\qquad\qquad m - l(c) \le i \le m \end{aligned} \tag{7}$$

where i and j are the index of the action and state, respectively, included in the step length $l(c)$, and m is the number of steps of the state-action trace leading the agent to the fail state. Since $p^N_{max_{ij}} = 1$ can not be defined in (4), we introduce a new notation 1^- as a value just less than and close to 1. It should be noted that if the failure probability of the action a_{f1} is 1^-, the action almost surely leads the agent to the fail state.

In the nondeterministic environment, the same action, which led the agent to the fail state in the previous trial, may not lead the agent to the fail state again. Setting the value of $p_{F_{ij}}$ to 1^- is a possible way to inhibit the corresponding action from the progress of the Q-learning in the nondeterministic environment. Thus, the failure probability of the selected action just before the fail state should be decreased as the trial goes on. For this purpose, the following decreasing scheme for $p^N_{max_{ij}}(k)$ is proposed:

$$p^N_{max_{ij}}(k) = \eta^{k-k_f} \tag{8}$$

where η is a constant between 0 and 1, k is the current trial number, and k_f is the failure trial number. In this paper, $p^N_{min_{ij}}$ is fixed as a constant value irrespective of N.

Figure 3 shows $p^N_{max_{ij}}(k)$ of (8) with respect to the trial k for the example depicted in Figure 2, where three failure trials exist. As the number of trials increases, the maximum failure probability $p^N_{max_{ij}}(k)$ of the action $(a_{fi}, i = 1,2,3)$ just before the fail state, decreases according to (8). However, if the same action that led the agent to the fail state is selected again, its maximum failure probability is set to 1^- (see the $(n+2)$th trial in Figure 3 (a)) and then it decreases again as the trial goes on.

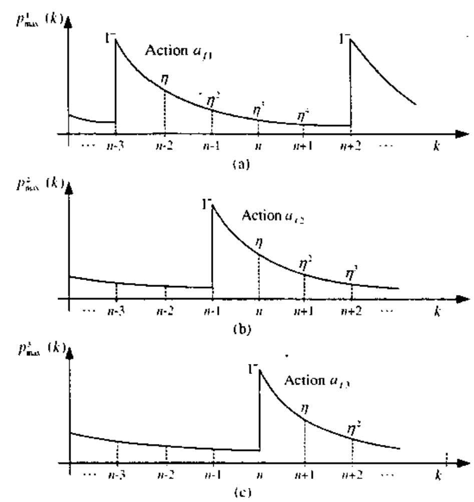


Figure 3: Maximum failure probability distribution for actions a_{f1}, a_{f2} and a_{f3}

When the agent reaches the fail state by the action a_{f1} in the $(n-3)$th trial, the maximum failure probability of the action a_{f1} is set to 1^- (Figure 3 (a)). As the trial goes on, $p^1_{max_{ij}}$ decreases by the factor of η. If in $(n-1)$th trial the agent arrives at the fail state by the action a_{f2}, then $p^2_{max_{ij}}$ becomes 1^- (Figure 3 (b)). In Figure 3 (c), when the agent enters into the fail state by a_{f3} in the nth trial, $p^3_{max_{ij}}$ becomes 1^-. At that moment the maximum failure probability of actions a_{f2} and a_{f1} becomes η and η^3, respectively. As the trial is proceeded, the $p^N_{max_{ij}}$ value decreases. If the agent reaches the fail state in the $(n+2)$th trial by a_{f1}, $p^1_{max_{ij}}$ becomes 1^- again (Figure 3 (a)).

When the agent reaches the fail state, the failure probabilities of the actions included in the step length $l(c)$ of the

procedure the two mode Q-learning
begin
  initialize $Q_{N_{ij}}$, $Q_{E_{ij}}$ and $Q_{T_{ij}}$
  $k \leftarrow 0$
  while (k < MAX_TRIAL) **do**
  begin
    $k \leftarrow k + 1$
    start from an initial state
    while (current state is not success state or fail state) **do**
    begin
      calculate $Q_{T_{ij}}$ (= $Q_{N_{ij}} + Q_{E_{ij}}$)
      select an action by $Q_{T_{ij}}$
      observe reward and next state
      update $Q_{N_{ij}}$
      if (next state is fail state) **then**
        clear all $p_{F_{ij}}$ except $p^N_{max_{ij}}(k)$
        calculate step length for $p_{F_{ij}}$
        apply new $p_{F_{ij}}$ to the actions within the step length
        calculate $p^N_{max_{ij}}(k)$
      else
        goto next state
        current state $\leftarrow$ next state
        if (current state is success state) **then**
          clear all $p_{F_{ij}}$ except $p^N_{max_{ij}}(k)$
    end
  end
end

Figure 4: Procedure the two mode Q-learning

state-action trace are cleared and assigned a new value calculated by (7). When the agent reaches the success state, all the failure probabilities of the state-action pairs are cleared to 0 except the $p^N_{max_{ij}}(k)$ value of a_{fi}.

Figure 4 summarizes the algorithm of the proposed two mode Q-learning.

In Q-learning, the agent learns the method of finding the optimal action in the state-action space. For finding the optimal action in a given state, the agent needs a lot of experiences in exploring the state-action space. In two mode Q-learning, the agent may have less failure experience as the trial goes on, compared to the past trials because the agent makes use of the previous failure experiences in the current trial. As a result of using the past failure experiences, the agent has more chance of not only having the failure experience but having more chances of traversing in the meaningful state-action space.

3 Simulation result

To investigate the performance of the two mode Q-learning in the nondeterministic environment, a robot soccer system was selected as a test bed. Figure 5 shows the playground and the soccer robot of the NaroSot category in the FIRA games. The goalkeeping ability of the goalie is used for comparing the performance between the conventional Q-learning and the two mode Q-learning. To apply them to the goalie robot, 324 ($2 \times 9 \times 2 \times 9$) states and 7 actions per state were defined in the following (see Figure 6):

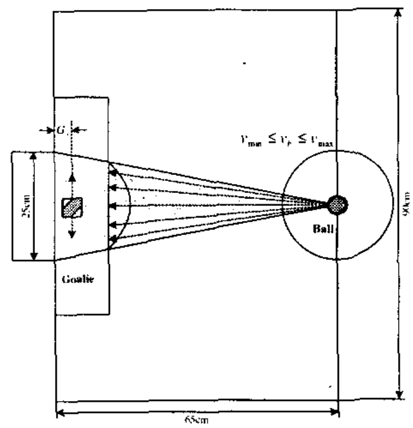


Figure 5: Goalie of the NaroSot

- States
  i) Flag of the goal in or not to the opponent goal post (2 states)
  ii) Distance in terms of the y-coordinate between the robot and the ball (9 states)
  iii) Sign of the distance in the y-coordinate between the robot and the ball (2 states)
  iv) x-coordinate of the ball position (9 states)
- Actions: Defined as the velocities of the robot, $\pm 20\ cm/s$, $\pm 10\ cm/s$, $\pm 4\ cm/s$, $0\ cm/s$ (7 actions)
- Reward
  i) Reward 1: In case of blocking the ball, $r = 100$
  ii) Reward 2: In case of not blocking the ball, $r = -100$

It should be noted that to block the ball, the goalie selects one of the 7 velocities as its action.

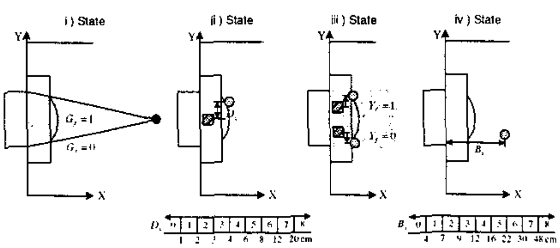


Figure 6: States for the goalie robot

The following kinematics of the two wheeled mobile robot was used for the goalie robot:

$$\begin{bmatrix} \dot{x}_c \\ \dot{y}_c \\ \dot{\theta}_c \end{bmatrix} = \begin{bmatrix} \cos\theta_c & 0 \\ \sin\theta_c & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} v \\ \omega \end{bmatrix} \tag{9}$$

where x_c and y_c are the x and y coordinate values of the robot center, θ_c is the heading angle of the robot, v is the translational velocity of the robot center, and ω is the rotational velocity with respect to the robot center. The following control law was used for the velocity control of the goalie:

$$\begin{aligned} dx &= G_o - x_c,\ dy = v \\ \theta_d &= \tan^{-1}(\frac{dy}{dx}) \\ \theta_e &= \theta_d - \theta_c \\ v_r &= v + k_p\theta_e + k_d\dot{\theta}_e \\ v_l &= v - k_p\theta_e - k_d\dot{\theta}_e \end{aligned} \tag{10}$$

where G_o is the offset (see Figure 5), which is given as the positive constant value, x_c is the x-coordinate value of the robot center, v is the translational velocity of the robot center. θ_e is the angle error, v_r and v_l are the right and left wheel velocities of the robot, respectively, and $\dot{\theta}_e$ is the time differential value of θ_e.

Considering the frictional force between the ball and the playground, the velocity of the ball was modelled as follows:

$$v_b = \begin{cases} v_b - C_{b_1}\frac{k}{D} & \text{if} \quad v_b \geq 20.0\ cm/s \\ v_b - C_{b_2}\frac{k}{D} & \text{else if}\ v_b \geq 8.0\ cm/s \\ v_b - C_{b_3}\frac{k}{D} & \text{otherwise} \end{cases} \tag{11}$$

where C_{b_1}, C_{b_2} and C_{b_3} are constant values, k is the trial number of the simulation, and D is the positive constant value. The velocity of the ball, v_b is bounded by $v_{min} \leq v_b \leq v_{max}$, and the angle of the ball, θ_b is restricted to $\theta_{min} \leq \theta_b \leq \theta_{max}$.

In the simulation, G_o=5, $v_{min} = 40cm/s$, $v_{max} = 70cm/s$, $\theta_{min} = 165°$, $\theta_{max} = 195°$, $C_{b_1} = 0.0012$, $C_{b_2} = 0.0006$, $C_{b_3} = 0.0003$, and $D = 1$ were used. Also, learning rate, $\alpha = 0.1$, discount rate, $\gamma = 0.9$, the minimum failure probability, $p^N_{min_{ij}} = 0.1$, the parameter of the maximum failure probability, $\eta = 0.9$, $1^- = 0.99$, and the constant value of step length, $c = 4$ were used.

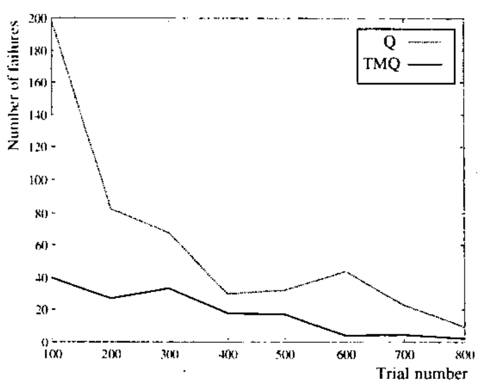


Figure 7: Result of goalkeeping ability of the goalie in the simulation

Figure 7 shows simulation results, where every points represent the average number of the failures of blocking the ball for 20 iterations (1 iteration = 800 trials). In this figure, Q and TMQ denote the conventional Q-learning and the two mode Q-learning, respectively. During the progress of the learning, goalie tests of blocking the ball were executed 1000 times at the end of every 100 trials. The number of the failures of blocking the ball was compared between the conventional Q-learning and the two mode Q-learning.

As the trial goes on, the number of failures of the goalie trained by the two mode Q-learning is less than that of the goalie by the conventional Q-learning. It means that the performance of the goalie trained by the two mode Q-learning is better than that of the goalie by the conventional Q-learning.

The result of the simulation of two mode Q-learning was implemented to the real soccer system. Figure 8 shows the NaroSot robot. The specification of the NaroSot robot is shown in table 1.

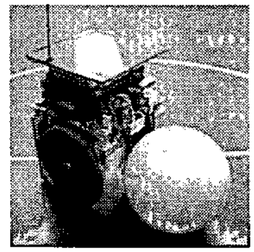


Figure 8: NaroSot robot

Size	$4cm \times 4cm \times 5.5cm$
Controller	ATmega163(ATmel)
Communication speed	19200 bps
Weight	130 g
Maximum velocity	$120cm/s$
Acceleration	$300cm/s^2$

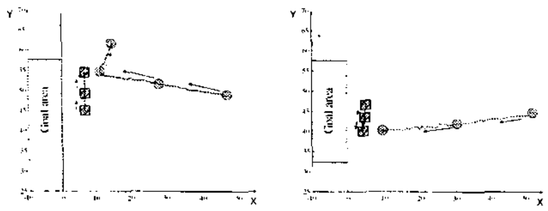
Table 1: Specification of the NaroSot robot

Table 2 shows the specification of the vision system used in the real experiment.

Host computer	Pentium III 800MHz
Camera resolution	640 pixel × 480 pixel
Image processing rate	60 Hz
Position error	Within 1.0 cm
Angle error	Within 5°

Table 2: Specification of the vision system

The Q values obtained by the simulation result of the two mode Q-learning were implemented to the NaroSot system. For implementing the Q values to the NaroSot system, the information of the robot and the ball was divided into the state as shown in figure 6 and the robot selected the action which had the maximum Q value in the state. At that moment, the position of the robot and the ball were shown in Figure 9.



(a) (b)

Figure 9: Experiment result of implementing the Q value obtained by the two mode Q-learning simulation to the NaroSot system

4 Summary and conclusion

This paper proposed a new two mode Q-learning by introducing a failure Q-value module into the conventional Q-learning. The failure Q value module applies the failure probability to the actions within a specific step length of the failure state-action trace, to calculate their failure Q value. Based on both normal Q value and failure Q value, an action is selected in the two mode Q-learning. As a result of using the past failure experience, the agent has more possibility of not having the failure experience again and will use the experience to explore in some meaningful fashion the state-action space. To investigate the effectiveness of the proposed two mode Q-learning in training a soccer agent to perform goalkeeping.

Bibliography

[1] L. P. Kaelbling and M. L. Littman and A. W. Moore, "Reinforcement Learning: A Survey," *Journal of Artificial Intelligence Research*, pp. 237-285, 1996.

[2] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, Bradford Books/MIT Press, 1998.

[3] G. A. Rummery, "Problem Solving with Reinforcement Learning," Ph.D thesis, Cambridge Unversity, England, 1995.

[4] M. N. Ahmadabadi and M. Asadpour, "Expertness Based Cooperative Q-Learning," *IEEE Trans. Systems, Man and Cybernetics*, Part B, vol 32, no. 1, pp. 66-76, 2002.

[5] C. J. C. H. Watkins and P. Dayan, "Q-learning,", *Machine Learning,*, vol. 8, pp. 279-292, 1992.

[6] M. A. Wiering, R. Salustowicz, and J. Schmidhuber, "Reinforcement Learning Soccer Teams with Incomplete World Models," *Autonomous Robots*, Special issue on Neural Networks for Robot Learning, vol 7. no. 1, pp. 77-88, 1999

[7] W. D. Smart, "Making Reinforcement Learning Work on Real Robots," Ph.D thesis, University of Brown, 2002.

[8] Y. Takahashi and M. Asada, "Behavior Acquistion by Multi-layered Reinforcement Learning," *Proc. of IEEE Int. Conf. Systems, Man, and Cybernetics*, vol. 6, pp. 716-721, 1999.

[9] O. Abul, F. Polat, and R. Alhajj, "Multiagent Reinforcement Learning Using Function Approximation,", *IEEE Trans. Systems, Man and Cybernetics*, Part C, vol 30, no. 4, pp. 485-497, 2000.

[10] D.-H. Kim, Y.-J. Kim, K.-C. Kim, J.-H. Kim, and P. Vadakkepat, "Vector Field Based Path Planninig and Petri-net Based Role Selection Mechanism with Q-learning for the Soccer Robot System," *Int. Journal of intelligent Automation and Soft Computing*, vol. 6, no. 1, pp. 75-88, 2000.

[11] K.-H. Park, Y.-J. Kim, and J.-H. Kim, "Modular Q-learning based multi-agent cooperation for robot soccer," *Robotics and Autonomous Systems*, vol. 31, no. 2, pp. 109-122, 2001.

[12] M. Asada, E. Uchibe, and K. Hosoda, "Cooperative behavior acquisition for mobile robots in dynamically changing real worlds via vision-based reinforcement learning and development," *Artificial Intelligence*, vol. 110, pp. 275-292, 1999.

[13] Y. Arai, T. Fujii, H. Asama, H. Kaetsu, and I. Endo, "Collision avoidance in multi-robot systems based on multi-layered reinforcement learning," *Robotics and Autonomous Systems*, vol. 29, no. 1, pp. 21-31, 1999

[14] S. P. Singh and R. S. Sutton, "Reinforcement learning with replacing eligibility traces," *Machine Learning*, vol. 22, pp. 123-158, 1996.

[15] J. Peng and R, J. Williams, "Incremental multi-step Q-learning," *Machine Learning*, vol. 22, pp. 283-290, 1996.

[16] M. A. Wiering, "Fast on-line Q(λ)," *Machine Learning*, vol. 33, no. 1, pp. 105-115, 1998.

[17] P. V. C. Caironi and M. Dorigo, "Training and Delay Reinforcements in Q-Learning Agents," *Int. Journal of intelligent systems*, vol. 12, no. 10, pp. 615-724, 1997.