Hide and Seek

107201535 陳羽暉
107201023 蔡沐霖
107201016 高文顥

PROBLEM FORMULATION AND SOLUTIONS

The goal of the project is using Q-learning to train two robots , one police and one thief , the police catch the thief and the thief escapes from the police.

The police and the thief are setting in a 2D plane with size of 30*30 , the position is denoted by coordinates (x,y) , $x,y \in N$ and $1 \le x,y \le 30$. There are 9 monitors evenly distributed at (5*i,5*j) where i,j = 1,3,5 , each monitor can observe 3*3 range ,and 2 outlets are sitting in fixed position (1,30) and (30,30).

The MDP of the police includes the state S=(x,y) , which is the current position of the police , the initial state $S_0=(x_0,y_0)$, which is the initial position of the police and it is random, the set of actions A={up,left,down,right,stay} , there will be no uncertainty so the transition probabilities of doing each action should be 1, if its action is stay , its arrest range will become 3*3.

The reward of doing an action is -0.1 if police catch the thief (at the same position) then it will get the reward 10 ; otherwise,if the thief arrives at the outlet then it will get the reward -10, the discount γ=0.9 .

The MDP of the thief is similar to the MDP of the police , but the distance of the initial position of the thief to the outlet will be farther than the police , and there are only four actions {up,left,down,right} for the thief. If the thief arrives at the outlet , then it will get the reward 10 ; otherwise,if it got caught(at the same position) ,it will get the reward -10.

For the monitor , both the police and the thief can know the situation of the monitor , but only the police can get the information(the situation of the thief) from the monitors.

When the police think the thief will go through a state(based on the monitor that finds the thief and the location of the outlet).

The features of the Q function of the police f1:the minimum distance to the two outlets;f2:the x of the location, f3:the y of the location;f4:the distance of the predicted location of the thief with itself.

The features of the Q function of the thief f1:the minimum distance to the two outlets;f2:the x of the location, f3:the y of the location;f4:the times that it passed the monitors.

$$Q(s,\ a)\ = \sum_i w_i f_i\,(s,a)$$

$$w_i \leftarrow w_i + \alpha\left[R(s) + \gamma \max_{a'} Q(s',a') - Q(s,a)\right]\frac{\partial Q(s,a)}{\partial w_i}$$

Q(s, a) is Q function at state s and do action a, γ is discount factor 0.9 and $\alpha$is learning rate 0.001 , w is the weighting vector.