Reinforcement Learning Applied in a Police Pursuit

Bernardo Eichler (77988)   Ricardo Sequeira (79750)

1. INTRODUCTION

Our project consists in four police cars pursuing a thief car in a particular city. The objective is to catch the thief, traping him between two cells from which he can't run away. The police cars respond to a police station that controls all the pursuit by registering the thief position when he is seen by one of the agents (since his position is unknown in the beggining of the problem).

2. AGENTS : POLICE CARS

Our main agents are the police cars because they are the entity to be studied in this problem. Our goal was to understand what is the best solution to catch a wanted car (in this case a thief car): should the police car go after the thief since the moment he sees it? Or should he do some barricade in the end of a street to trap him?

2.1 Sensors

A police car knows his position and where can he head to (his possible directions). He even recognizes the city boundaries from a distance of 3 cells. The same distance from which he can identify the thief. The rest of the information that he gets is from the police station that he is connected and which informs him if the thief position is known or not, and if it is, where is him.

2.2 Actuators

The police car can move around the map in any legal direction (legal directions are embeded in the map cells) to another cell that isn't occupied by another car. There is this special case when the police gets to know a thief position (through the police station communication) and can choose a illegal direction (a direction that isn't embeded in the cell but allows him to get to another road cell inside of the map).

3. ENVIRONMENT: CITY AND THIEF

We wanted to have a rich environment so we envisioned a city with civil cars driving to given legal directions (embeded in map cells) leaving and entering again in the map (normal traffic). However the difficulty of our problem given those conditions was too high so we cut our vision for the city and we ended with a close simple map (police and thief don't leave the city because the city is closed for the pursuit) without civil cars (but the source code for them is there).

The thief is implemented as a car just like the police ones, with similar sensors and actuators. That said, we can say that the thief movement is completely random and only changed by the position of other cars (in this case, there are only other police cars). This last information will be important for our analysis.

The thief was intented to be searching for a garage which would grant him a way out of the city and be succeded in the pursuit. But once more, this proved to be a big complexity in our problem so we cut it and made the thief position in the map random as already said. The thief has also the capacity of choosing (randomly) illegal directions when trapped by the police.

4. REINFORCEMENT LEARNING

Since the beginning that our intention was to use reinforcement learning in the decision making algorithm, in particular, using the Q learning technique to calculate the quality of every police car move. The common rewards for every agent is assured by the police station which distributes the reward value.

4.1 Reward Function

The reward function that we implemented gives a maximum of 100 if the police car has directly arrested the thief, gives 70 if the police car is moving to a cell in the thief direction (if his position is known) and finally, give a bonus reward for how recently the cell of his next position was visited: if the cell was recently visited the reward is minimal here, if it was never visited (or in the last 10 steps) the reward is maximum.

4.2 Adjustements

The direction decision during the train iterations is not completely random because we made the police cars to choose the best action (calculated until there) half of the times and choose a random action the other half (to not get stuck in local maximum).

4.3 Problems

After the training with every number of iterations (we tried with lots of them so that was not the problem) we concluded that we have a problem with our Q. It's possible that is due to our reward function being too simple or maybe because the thief position is completely random according to the position of the police cars. We tried to fix our Q but we failed in the end and we couldn't pass to the next step.

5. THE NEXT STEP

We wanted to, after fixing the decision making algorithm, implement the police station as a unique agent who would decide to call more police cars (at a financial cost) given the time that a given pursuit is taking. With that, we wanted to study the problem of how many police cars are needed and the perfect time for a pursuit. We wanted also to give the police cars a personality: a more selfish one (who wants to be rewarded for catching the thief) or a more collaborative one (who wants to help catch the thief and reward everyone).

6. CONCLUSIONS

Unfortunately, we didn't make to any conclusion because of the problems in the implementation of the Q learning technique.