Study and Application of Reinforcement Learning in Cooperative Strategy of the Robot Soccer Based on BDI Model

Guo Qi and Wu Bo-ying
Department of Mathematics, Harbin Institute of Technology, Harbin 150001, China

Abstract: The dynamic cooperation model of multi-Agent is formed by combining reinforcement learning with BDI model. In this model, the concept of the individual optimization loses its meaning, because the repayment of each Agent does not only depend on itself but also on the choice of other Agents. All Agents can pursue a common optimum solution and try to realize the united intention as a whole to a maximum limit. The robot moves to its goal, depending on the present positions of the other robots that cooperate with it and the present position of the ball. One of these robots cooperating with it is controlled to move by man with a joystick. In this way, Agent can be ensured to search for each state-action as frequently as possible when it carries on choosing movements, so as to shorten the time of searching for the movement space so that the convergence speed of reinforcement learning can be improved. The validity of the proposed cooperative strategy for the robot soccer has been proved by combining theoretical analysis with simulation robot soccer match (11vs11).

Keywords: robot soccer; BDI model; Reinforcement Learning; joystick; human-control robot.

1. Introduction

The study of the robot soccer simulation focuses on the advanced functions adopted by a team, including the current hot issues of artificial intelligence, such as multi-Agent cooperation, real-time reasoning—planning—decision-making, the machine learning and obtainment of strategies in the uncertain dynamic environment. At present, some study technologies of the entrant teams have been realized, such as genetic algorithm [1], supervision learning and reinforcement learning. Reinforcement learning is used most frequently [2, 3, 4, 5]. Karlsruhe Brainstormer (a team in Germany) regarded reinforcement learning as their main research direction and had made the outstanding achievement: they won the runner-up of the simulation groups in 2000 RoboCup. Their goal is that the intellectual body can communicate with the environment independently and study various skills and strategies by only being told to finish scoring [6].

We have also done some research on the application of reinforcement learning to the robot soccer. We mainly study the application of the reinforcement learning to the multi-Agent cooperation [7], but not to the rudimentary behavior of an individual Agent (e.g. kick)[8].

In the research of multi-Agent, people have proposed the thinking model of multi-Agent—BDI model. This model includes the connection and activity relation between thinking property of multi-Agent and Agents. Among them, basic thinking property of multi-Agent is composed of belief, desire and intention. In BDI model of multi-Agent, belief is to describe the estimation of Agent at the present environment and the behavior it may adopt; desire is used to describe that Agent has a taste for the environment and behavior that it may adopt in the future; intention is used to describe that the Agent has made a promise in order to reach a certain goal.

Though multi-Agent represents different fields, and has its own local interests and local targets, while solving the problem in coordination, it should solve the problem in coordination and in an orderly manner under the restraint of the common planning. That is to say, the course of local solving must be adjusted in time according to the change of the external environment, in order to reach the interaction, dynamics and timeliness of the solving course in coordination. From the point of concept, BDI model based on the logical method has realized rational reasoning, while reinforcement learning has realized decision-making through optimization utility. From the point of technology, the logical method that adopts symbols to reason is unable to make utility best, and reinforcement learning with the numerical analysis has also ignored reasoning. As to a multi-Agent system of the dynamic environment, it is necessary to reason for the environmental message and to make a reasonable decision which may bring the maximum profit through learning.

In this paper, we combine reinforcement learning with BDI model [9, 10], so as to form a dynamic cooperative model of multi-Agent. In this model, the concept of the individual optimization loses its meaning, because the repayment of each Agent does not only depend on itself.
but also on the choice of other Agents. All Agents can pursue a common optimum solution and try to realize the united intention as a whole to a maximum limit. The robot moves to its goal, depending on the present positions of other robots that cooperate with it and the present positions of ball. One of these robots cooperating with it is controlled to move by man with a joystick. In this way, Agent can be ensured to search for each state-action as frequently as possible when it carries on choosing movements, so as to shorten the time of searching for the movement space so that the convergence speed of reinforcement learning can be improved.

2. The Cooperative Strategy of the Robot Soccer in Match

2.1. The Attack Strategy

The research on the simulation robot soccer match (11 vs 11) is presented in this paper. The soccer field is divided as follows: The halfway line is made as the boundary; on the left of the halfway line, including the opposite goal areas is the fore court; on the right of the halfway line, including our goal areas is the back court. When the ball is at the fore court, the attack strategy is adopted; when the ball is at the back court, we adopt the defensive strategy. The fore court is divided into 8 areas. The division of the area and the coordinate orientation of the court are shown in Fig.1.

The attack strategy: The members for attacking are divided into two groups—the shooting group and the mend-shooting group. Certain distance is kept between the members of two groups. All the members of two groups are before the opposite goal area (see Fig. 2). The shooting group: It is made up of 7 members. Its main task is to shoot the ball into the opposite goal. The mend-shooting group: It is made up of 3 members, and its main task is to shoot the ball into the opposite goal again, which isn’t shot into by the shooting group.

There are two kinds of possible situations: (1) The member of the shooting group has shot the ball, but the ball comes out from the opposite goal corner to the position where the members of the mend-shooting group stand; (2) The member of shooting group hasn’t shot the ball into the opposite goal, but shot the ball from another side in front of the opposite goal area to the position where the members of the mend-shooting group stand.

The strategy of the shooting group: The positions of the shooting group members follow the present position of the ball closely. The one who is closest to the ball will shoot.

The strategy of the mend-shooting group: The positions of the mend-shooting group members are correlative with the position of ball and other members’ positions in the same group. The one whose distance from the ball is less than 100 will shoot.

The strategy of the mend-shooting group: The positions of the mend-shooting group members are correlative with the position of ball and other members’ positions in the same group. The one whose distance from the ball is less than 100 will shoot.

The strategy of the mend-shooting group is simpler, while the strategy of the mend-shooting group, i.e. cooperative strategy, is more complicated. The reinforcement learning based on BDI model proposed in this paper mainly focuses on the cooperative strategy of the three members in the mend-shooting group.

2.2. The Study Task

The main task of the three members in the mend-shooting group is to determine the correct position of themselves at present according to the present position of the ball and other members in the same group and then shoot the ball to the opposite goal accurately when the ball comes. The task of the members in the mend-shooting group includes two steps:

The first step: According to the positions of the ball and the other members in the same group, they look for their own proper position;

The second step: Shoot the ball into the opposite goal.

The first step of the task determines whether the second step of the task will succeed or not: if the position where the members stand is proper, they will receive the ball and shoot the ball into the opposite goal, otherwise they can’t receive the ball, or after receiving the ball, they can not shoot the ball into the opposite goal in a right
Suppose the different, but in each of these areas, the three members all the cooperative strategies of the three members are space can be finished. When the ball is in each of 8 areas, divided areas of the fore court, the dispersion of state long as we make the position of ball dispersed in the 8 areas. This position of the member A is decided by present position of the ball, and the present positions of the other two members are not only related to the position of the ball but also to the other members’ positions in the same group. In this way, three members could cooperate closely and pursue the common optimum solution − if the relative position of the three members is the best, the scoring rate will be the highest.

Setting the position coordinates of three members in the mend-shooting group are separately:

\[
\begin{align*}
    x_a &= l + x_{ball} \\
    y_a &= h + y_{ball} \\
    x_b &= a_1 x_a + b_1 x_{ball} + c_1 x_c \\
    y_b &= a_1 y_a + b_1 y_{ball} + c_1 y_c \\
    x_c &= a_2 x_a + b_2 x_{ball} + c_2 x_b \\
    y_c &= a_2 y_a + b_2 y_{ball} + c_2 y_b
\end{align*}
\]

Where, \(l, h, a_1, b_1, c_1, a_2, b_2, c_2, a_1, b_1, c_1, a_2, b_2, c_2\) are separately the definite constants called control parameters. Every member has a input variable—the present position of the ball. These 14 control parameters can determine the correct position of the three members in the dynamic environment, so these 14 control parameters are decided by reinforcement learning studied in the paper.

### 2.3. The Division of The State Space

We know that the environment of the robot soccer is continuous. For example, the input includes the present position (two-dimentional vector) of the ball and robots, and it is continuous; the output is a sequence of movements and it is continuous too. But reinforcement learning is suitable for the solution of dispersed space. In this paper, so long as the present position of the ball is known, the positions of three members \(P_a, P_b, P_c\) can be worked out by the 14 control parameters. In this way, as long as we make the position of ball dispersed in the 8 divided areas of the fore court, the dispersion of state space can be finished. When the ball is in each of 8 areas, the cooperative strategies of the three members are different, but in each of these areas, the three members all have a best position, which will make the rate of shooting the highest and bring the maximum repayment. Because the position of the ball is at random, it can not evenly pass each area or the even dispersed positions of each area. Therefore, it is very difficult to search for the optimum solution. Because of this, the movement of the member A is controlled with the joystick in this paper, making it able to keep proper distance from the ball which moves at random so as to obtain the best rate of scoring. Because the other two members’ positions correlate with the positions of the ball and the other members in the same group, if the best movement of the member A is set with the joystick − \(l, h\) are known, then the best movements of the member B and the member C are determined correspondingly too, and 14 control parameters \(l, h, a_1, b_1, c_1, a_2, b_2, c_2, a'_1, b'_1, c'_1, a'_2, b'_2, c'_2\) are determined. So the speed of searching for the state space can be improved greatly with the use of the joystick.

### 3. BDI Model Based on Reinforcement Learning and The Realization of BDI Model

#### 3.1. The basic principle of Reinforcement Learning

Reinforcement Learning is also called encourage learning. It is a method without supervision. It is the action of Agent to communicate with the dynamic environment by the method of trial-and-error and to study the correct action independently. The Agent utilizes the already having knowledge to try various movements constantly and to revise its own knowledge and strategy through the rewards and punishments from the environment, so as to find or approach the strategy that can get the maximum reward. The standard model of reinforcement learning composed is as follows:

Strategy 1: \(S \rightarrow A\), mapping from the environmental state to action, function or a corresponding table;

Reward function \(R(s, a): S \times A \rightarrow R\) The reward after a movement is carried out;

Reward function for a long time \(V(s): S \rightarrow V\) The reward after a series of movements are carried out;

Environmental model \(T\) (optional): \(S \times A \rightarrow R \times R_0\)

#### 3.2. The BDI Model Based on Reinforcement Learning

We combine reinforcement learning with BDI model in order to solve the decision-making problem of multi-Agent. We define it as multiple members:

Definition 1. In \([S, G, A, B, D, I, \lambda]\), S is the dispersed state space, G is the goal solved in coordination, A is a set of multi-Agent, B is a belief set of multi-Agent, D is a desire set of multi-Agent, I is an intention set of multi-Agent, and \(\lambda=\{\lambda_1, \lambda_2, ..., \lambda_n\}\) is a set of value coefficient when multi-Agent carries on solving the problem, \(\lambda \in [0, 1]\).

Generally speaking, each Agent solves the problem locally according to their own local planning, and does not consider the movement planning of other Agents, so there must be conflicts of intention between Agents, which makes the equation below established:

\[
\sum_{i=1}^{n} \lambda_i = 0
\]
The internal motive force of the multi-Agent system is the intention, and the instability comes mainly from the intention conflicts. According to the theory of dynamics, systematic movement is always towards the stable equilibrium attitude and far away from the unstable equilibrium attitude. To the Agent in system, whether in conflict or cooperation, it always reaches a certain equalization point or steady point finally. If we divide multi-Agent system into two parts: Agents whose intention have conflict and Agents whose intention have no conflict, then the system can be regarded as the gaming system of a pair of matrixes. According to Nash theorem, there must be a balance of mixed strategies in the gaming system of a pair of matrixes. According to the theory of dynamics, the united intention as a whole to the maximum extent.

(I) The punishment rule: Using the Q-learning algorithm with Q value as the appraisal value of the state-movement, the rule of rewards and punishments is adopted and the task succeeds, then \( r = 1 \) (corresponding to the abstract movement \( a[P_1, P_2, \text{success}] \));

2. The member moves from the position \( P_1 \) to the position \( P_2 \). The follow-up task (shooter) can be adopted and the task fails, then \( r = 0 \) (corresponding to the abstract movement \( a[P_1, P_2, \text{failure}] \)).

3. The member moves from the position \( P_1 \) to the position \( P_2 \). The follow-up task can’t be adopted (the distance between the member and the ball is greater than 100), then \( r = -1 \) (corresponding to the abstract movement \( a[P_1, P_2, 0] \)).

(II) Determine the environment and action:

Set of the state: \( S = S_1 \cup S_0 \), the state in \( S_1 \) is the position of the ball \( P = (x_{ball}, y_{ball}) \), and the positions of three members \( P_k \) (\( k = 1, 2, 3 \)); the state in \( S_0 \) is the position of the ball and the three members, in which the follow-up task can be adopted and the task succeeds.

Action: \( A \subseteq S \), the state (position) that can be reached from a kind of state (position) after carrying out a basic order: position (robot, cpoint\((P_i, P_j)\)).

(III) Determine the parameter (immediate reward) \( r \):

1. The follow-up task can be adopted, and the task succeeds, then \( r = 1 \);
2. The follow-up task can be adopted, and the task fails, then \( r = 0 \);
3. The follow-up task can’t be adopted, then \( r = -1 \).

(IV) The algorithm of searching online:

1. Search for \( S \), search for \( S_0 \) from \( S \), \( S_0 \subset S \);
2. Determine the action \( a \) from \( S_0 \), \( a \) represents the action with maximum Q value

\[
Q(S_0, a) = \max_a Q(S, a) 
\]

3. The goal position of the action is determined by 14 control parameters in (1). Send the basic order that action \( a \) needs: \( a[P_1, P_2] \rightarrow \) position (robot, cpoint \((P_i, P_j)\))

3.3. Abstract Movements

The set of movements that can be chosen is described by an abstract movement. This abstract movement is:

\[ a[P_1, P_2] = P_1 \rightarrow P_2 \]

It means that a member can move from the position \( P_1 \) to the position \( P_2 \). It is realized by the function in programming: position (robot, cpoint\((P_i, P_j)\)), the first parameter of the function “robot” means the robot controlled; the second parameter \( P_1 \) is the original position of the robot; the third parameter \( P_2 \) is the goal position of the robot. There are also three special abstract movements:

\[ a[P_1, P_2, \text{success}] \]: This abstract movement shows the member moves from the position \( P_1 \) to the position \( P_2 \). The follow-up task can be adopted and the task succeeds.

\[ a[P_1, P_2, 0] \]: It shows that the member moves from the position \( P_1 \) to the position \( P_2 \). The follow-up task can’t be adopted.

\[ a[P_1, P_2, \text{failure}] \]: It shows that the member moves from the position \( P_1 \) to the position \( P_2 \). The follow-up task can be adopted, but the task fails.

3.4. The Study Scheme

(I) The punishment rule: Using the Q-learning algorithm with Q value as the appraisal value of the state-movement, the rule of rewards and punishments is designed as follows:

1. The member moves from the position \( P_1 \) to the position \( P_2 \). The follow-up task (shooter) can be adopted and the task succeeds, then \( r = 1 \) (corresponding to the abstract movement \( a[P_1, P_2, \text{success}] \));

2. The member moves from the position \( P_1 \) to the position \( P_2 \). The follow-up task (shooter) can be adopted and the task fails, then \( r = 0 \) (corresponding to the abstract movement \( a[P_1, P_2, \text{failure}] \)).

3. The member moves from the position \( P_1 \) to the position \( P_2 \). The follow-up task can’t be adopted (the distance between the member and the ball is greater than 100), then \( r = -1 \) (corresponding to the abstract movement \( a[P_1, P_2, 0] \)).

\[ (II) \text{Determine the environment and action:} \]

Set of the state: \( S = S_1 \cup S_0 \), the state in \( S_1 \) is the position of the ball \( P = (x_{ball}, y_{ball}) \), and the positions of three members \( P_k \) (\( k = 1, 2, 3 \)); the state in \( S_0 \) is the position of the ball and the three members, in which the follow-up task can be adopted and the task succeeds.

Action: \( A \subseteq S \), the state (position) that can be reached from a kind of state (position) after carrying out a basic order: position (robot, cpoint\((P_i, P_j)\)).

(III) Determine the parameter (immediate reward) \( r \):

1. The follow-up task can be adopted, and the task succeeds, then \( r = 1 \);
2. The follow-up task can be adopted, and the task fails, then \( r = 0 \);
3. The follow-up task can’t be adopted, then \( r = -1 \).

(IV) The algorithm of searching online:

1. Search for \( S \), search for \( S_0 \) from \( S \), \( S_0 \subset S \);
2. Determine the action \( a \) from \( S_0 \), \( a \) represents the action with maximum Q value

\[
Q(S_0, a) = \max_a Q(S, a) 
\]

3. The goal position of the action is determined by 14 control parameters in (1). Send the basic order that action \( a \) needs: \( a[P_1, P_2] \rightarrow \) position (robot, cpoint \((P_i, P_j)\))

3.5. The Realization of the Model

The realization steps of the algorithm:

To each \( S \), \( a \), where \( a = (a', a^2, a^3) \), initialize \( \hat{Q}(S, a', a^2, a^3) \) into 0, observe the present state of \( S \), and repeat this process all the time:

- Choose a movement \( a \) and carry out it
- Receive the immediate reward \( r \)
- Observe the new state \( S' \)
- Update \( \hat{Q}(S, a', a^2, a^3) \) according to

\[
\hat{Q}_n(S, a', a^2, a^3) \leftarrow (1 - \alpha_n) \hat{Q}_{n-1}(S, a', a^2, a^3) + \alpha_n (r + \gamma \max_a \hat{Q}_{n-1}(S', a', a^2, a^3))
\]

Where, \( \alpha_n = \frac{1}{1 + \text{visit}_S(S, a', a^2, a^3)} \), \( r = \frac{1}{3} \sum_{i=1}^{3} r_i \),

conversion rate \( \gamma = 0.9 \), and \( a \) is a renewed couple of state-movement in the \( n \)-th circulation, and \( \text{visit}_S(S, a', a^2, a^3) \) is the total number of times of this
couples of state-movement visited in these \( n \) times of circulation (include the \( n^{th} \) times of circulation). The mark \( i(1 \leq i \leq 3) \) of \( a' \) and \( a' \) represents serial number of the three members.

4. Simulation

In view of the simulation robot soccer match (11 vs 11), the match strategy proposed in this paper based on reinforcement learning has been demonstrated in this experiment. In the 8 areas that are divided in the court, the optimum solutions of 14 control parameters are searched on-line. In order to avoid the influence of random factors on the experimental result, the experiments of many sessions have been carried out. When the optimum solution tends towards stability, 14 control parameters are determined, as shown in Table 1. Then with the 14 control parameters decided, the match is carried on based on the attack strategy proposed in this paper.

We have adopted two kinds of attack strategies to carry on the match:

The first is the attack strategy based on reinforcement learning proposed in this paper, which adopts the 14 control parameters in Table 1 to determine the positions of three members in the mend-shooting group, but the abscissa positions of the three members must meet the conditions: when \( x_i < 185 \), then \( x_i = 185 \) (\( i = a, b, c \)).

The second is the attack strategy without machine learning, which is adopted as the strategy (similarity 11 vs 11) of our team in the fifth national robot soccer championship match (won over Nanjing University of Technology, defeated by Chengdu University of Technology). While adopting the second kind of attack strategy to carry on the match, in each of the 8 areas of the court, each of the three members’ position correlates with the position of the ball, and has nothing to do with the positions of the other members in the same group. The position relations of the members and the ball are determined by experience, not by the method of machine learning. The position coordinate relation between the three members and the ball in the 8 areas are shown in Table 2.

| Area | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|------|---|---|---|---|---|---|---|---|
| 1    | 10.0 | 50.0 | -20.0 | 10.0 | 50.0 | -20.0 | 10.0 | 50.0 | -20.0 |
| 2    | 100.0 | 10.0 | -20.0 | 100.0 | 10.0 | -20.0 | 100.0 | 10.0 | -20.0 |
| 3    | -1.0 | 0.0 | 0.0 | -1.0 | 0.0 | 0.0 | -1.0 | 0.0 | 0.0 |
| 4    | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| 5    | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 6    | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 |
| 7    | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| 8    | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 |

Table 1. The 14 control parameters determined in the 8 areas.

The same rival is chosen (the rival’s strategy does not change) while these two kinds of attack strategies are adopted in the match. Each kind of attack strategy has been carried on in 10 matches. The experimental results of these two kinds of attack strategies adopted in matches are shown as Fig. 3 and Fig. 4. It is shown that, the score of the soccer team based on the dynamic cooperation model is obviously higher than that of the soccer team which hasn’t adopted reinforcement learning in the attack strategy.
5. Conclusion

Reinforcement learning is combined with BDI thinking model in this paper and the dynamic cooperative model is formed in view of multi-intelligent robots. In this model, the concept of the individual optimization has lost its meaning. All intelligent robots can pursue a common optimum solution, and try to realize the united intention as a whole to a maximum limit. Because the space searched for reinforcement learning is quite large, in order to accelerate the optimum solution convergence, one of the robots that cooperate with each other is controlled by man with a joystick. In this way, the intelligent robot can be ensured to search for each a couple of state-action as frequently as possible when it carries on choosing movements, so as to shorten the time of searching for the movement space so that the convergence speed of reinforcement learning can be improved. In the simulation robot soccer match (11 vs 11), the validity of the cooperative strategy proposed in this paper has been proved.

8. References

[1] Yin Xiang, Huang Zhan-xiang. Learning to Kick: Application of Reinforcement Learning Simulant Robotic Soccer, Journal of Soochow University (Engineering Science Edition), Vol.22, No.4, Aug. 2002, pp. 26-32

[2] Neeraj Arora and Sandip Sen. Resolving social dilemmas using genetic algorithms. In Adaption Coevolution and Learning in Multiagent Systems: Papers from the 1996 AAAI Spring Symposium pages 1-5, Menlo Park, CA, March 1996, AAAI Press. AAAI Technical Report SS-96-01 (pp. 188-191)

[3] DUAN Yong; XU Xin-he. Reinforcement learning based on FNN and its application in robot navigation. Control and Decision, 2007, Vol. 22, No.05, pp. 525-534

[4] Handa, Hisashi. Evolutionary Computation on Multitask Reinforcement Learning Problems. Networking, Sensing and Control, 2007 IEEE International Conference on 15-17 April 2007 Page(s): 685 – 688

[5] Wiering, M.A.; van Hasselt, H.; Two Novel On-policy Reinforcement Learning Algorithms based on TD(ýý)-methods. Approximate Dynamic Programming and Reinforcement Learning, 2007. ADPRL 2007. IEEE International Symposium on 1-5 April 2007 Page(s): 280 - 287

[6] Y. Takahashi and M. Asada, “Vision-Guided Behavior Acquisition of a Mobile Robot by Multi-Layered Reinforcement Learning”, in Proc. Of the 2000 IEEE/RSJ IROS, pp. 395-402, 2000

[7] Martin Riedmiller, Artur Merke, David Meier, Andreas Hoffmann, Alex Sinner, Ortwin Thate, Ralf Ehrmann. Karlsruhe Brain-stormers—a reinforcement learning learning approach to robotic soccer. RoboCup-00: Robot Soccer World Cup IV, LNCS, Springer

[8] A S Ral, G Murray. Multigragentmental-state recognition and its application to air-combat modelling. In: V R Lessered. Proc of the 13th Int’l Workshop on Distributed Artificial Intelligence. Seattle, WA, 1994. 283-304

[9] HU Shan-li, SHI Chun-yi, Agent-BDI Logic, Journal of Software, 2000, 11(10):1353 ~1360

[10] LU Jun, WANG Ya-dong, WANG Xiao-long, Research and Improvement of the Interpreter of BDI Agent, Journal of Software, 2000, 11(8):1118~1125