# Gemini in RoboCup-2002

Masayuki Ohta

Cyber Assist Research Center,
National Institution of Advanced Industrial Science and Technology,
AIST Tokyo Waterfront 2-41-6 Aomi Koto-ku Tokyo 135-0064 Japan
ohta@carc.aist.go.jp

**Abstract.** We implemented Gemini to research a lot of aspects of machine learning. Recently, we are focusing on reinforcement learning in multi-agent environment, and investigate the feature of direct reward and indirect reward. We did learning experiment in which agents acquire an effective positioning policy with Gemini. The empirical results show that direct reward takes effect faster and helps obtaining individuality. On the contrary, indirect reward takes effect slower, but agents tend to form a group. And now we are searching for an effective way to use both direct and indirect reward.

## 1 Introduction

Learning cooperative behaviors in multi-agent environment is an important topic in the research field of distributed artificial intelligence. We implemented "Gemini", which is a client program run on SoccerServer[3], to investigate various kinds of machine learning methods. Our major interest is about acquiring effective cooperation policy in noisy environment without direct communication. In this time, we are trying to solve a credit assignment problem of reinforcement learning with mixing two kinds of reward. In this paper, we explain Gemini's action rules briefly, and then show our experiments letting agents learn their positioning policy.

## 2 Overview of the Architecture

In the basic parts, Gemini is almost the same as Gemini in RoboCup 2001[4]. Each player behaves as about the following simple rules.

- if (lost sight of the ball) look for the ball
- else if(ball is in kickable area) shoot or pass or clear (hand-coded)
- else if(nearest from the ball) chase the ball
- else go back to the base position and trace the ball

The base position is recalculated every five seconds. They divide the soccer field into grid, then calculate the expected discounted reward in the case that their base position is at the center of the cell with neural network. The inputs of the neural network are x-coordinate, y-coordinate, distance to the closest teammate and distance to the closest opponent of the cell. The center of the cell whose expected reward is the best becomes the next base position.

## 3 Research Issues

When we apply reinforcement learning to acquire cooperative behaviors in multi-agent environment, deciding who should be praised is one of a serious problem. If we praise only directly contributing agent, they tend not to learn cooperative behaviors. On the other hand, if we praise all agents when some agents achieved their common goal, there is some possibility of reinforcing bad policies. We call these two kinds of reward "direct reward" and "indirect reward". We think it is useful to investigate the feature of these reward and use mixture of them in proper ratio.

To confirm this, we did learning experiment to acquire an effective positioning policy with Gemini[5]. We gave direct reward when they kicked the ball by themselves, and indirect reward when they saw a teammate kicking the ball. We adopted Monte Carlo method to estimate the expected discounted reward, because of its robustness against violations of the Markov property [6]. In the experiment, we considered an evolution of learning team against fixed positioning team, in which they completely have the same ability besides positioning. The result with only direct reward is Fig.1, and the result with only indirect reward is Fig.1, in which x-axis indicates the time scale by a minute (= 600 step), and y-axis indicates the score of learning team in a minute. In this experiment, we could see the following results.

- Direct reward affects learning very quickly, and indirect reward affects learning relatively slow.
- Both agents showed almost the same effectivity finally, but they acquired totally different policy.
- Agents learned with direct reward took distributed formation, and the agents learned with indirect reward took crowd formation.

Considering this result, when we use both of these reward together, we can expect that the score during the learning can be improved, if direct reward is used in a big ratio at first, and if the ratio is reduced along the time. Then we compared the following two conditions of rewarding.

- **Fix the mixture ratio**: Mix direct reward and indirect reward half and half. The ratio is never changed.
- **Change the mixture ratio**: Changed the ratio of direct reward from 20% (in the begining) to 80% (at 1000 minuters).

Fig.?? and Fig.4 show the learning curve with both rewarding method respectively. This result shows that, simple mixture does not increase the efficiency of the learning nor the speed of learning. It is expected that the reason is because two kinds of reward conflict each other, in this case. However, when we compare the both mixture, the method changing the mixture ratio could find much better policy. So with this experiment we can say that, if there are more than two kind of reward and if it can divide into direct reward and indirect reward, changing mixture ratio along the time is an effective solution. But, before start learning

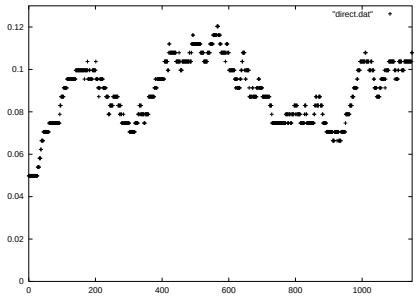


**Fig. 1.** Score of the learning team reinforced only by direct reward against a team with fixed policy

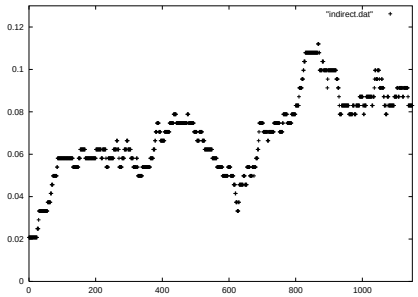


**Fig. 2.** Score of the learning team reinforced only by indirect reward against a team with fixed policy

we should know when the learning converge about, this can be a disadvantage of this method.

## 4 Related Works

Miyazaki[2] also classify the reward for reinforcement learning into direct reward and indirect reward. Their definitions are very similar to our definition, but different to some extent. They showed necessary and sufficient condition of direct and indirect reward ratio to preserve the rationality. In their research, they are interested in the fixed ratio, but we are focusing on the change along the time.

Observational Reinforcement Learning[1] solved the problem that exploration causes bad effect on the other agents' learning in multi-agent environment. They also treat a positioning task on SoccerServer. Their solution is to use a reward which reinforce the position where the agents think it is good, by their perception information. They used this reward mixed with the traditional method (direct reward in this paper). Using this rewarding method, agents can have almost the same effect as the exploration, even if it reduce their exploration ratio. This method is very useful, but it is a kind of supervised learning, and agents need to know some candidate answers beforehand. It is also interesting that, the agents reinforced by observational reward have similar feature to the agents reinforced by indirect reward.

## 5 Conclusion

We are trying to solve a credit assignment problem of reinforcement learning in multi-agent environment with Gemini. In our approach, we classified the reward into direct reward and indirect reward, and adjust the mixture ratio to the effective ratio. We did learning experiment in which agents learn an effective

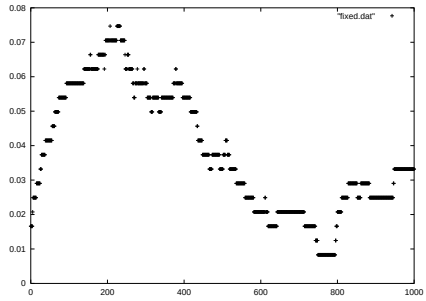


**Fig. 3.** The mixture ratio is fixed

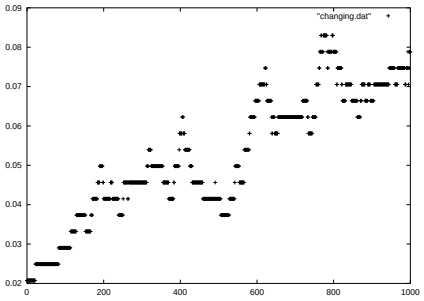


**Fig. 4.** The mixture ratio is changed

positioning policy in soccer simulation. An empirical results show that a simple mixture of direct reward and indirect reward does not increase the efficiency of the learning nor the speed of learning. But, if there are already these two kinds of reward, changing the mixture ratio is possibly effective.

## References

1. T. Andou. Andhill-98: A robocup team which reinforces positioning with observation. In M. Asada and H. Kitano, editors, *RoboCup-98: Robot Soccer World Cup II*, pages 338–345. Springer Verlag, 1998.
2. K. Miyazaki and S. Kobayashi. Rationality of reward sharing in multi-agent reinforcement learning. In *Second Pacific Rim International Workshop on Multi-Agents*, pages 111–125, 1999.
3. I. Noda, H. Matsubara, K. Hiraki, and I. Frank. Soccer server: A tool for research on multiagent systems. In *Applied Artificial Intelligence*, volume 12, pages 233–250, 1998.
4. M. Ohta. Gemini in robocup-2001. In *RoboCup 2001: Robot Soccer World Cup V*. Springer Verlag, 2001. In appear.
5. M. Ohta. Mixture of direct reward and indirect reward in multi-agent reinforcement learning. In *Proceedings of The Seventh Pacific Rim International Conference on Artificial Intelligence*, 2002. submitted.
6. R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.