Identification of opponent's strategy by active perception

Yoichi Motomura
Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba
Ibaraki, 305, JAPAN
motomura@etl.go.jp

Shotaro Akaho
Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba
Ibaraki, 305, JAPAN
akaho@etl.go.jp

Hideki Asoh
Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba
Ibaraki, 305, JAPAN
asoh@etl.go.jp

Abstract

In RoboCup competition, to realize flexible adaptative ability is an important issue as with a standard problem in AI. For this purpose, identification of an opponent's strategy is studied in this paper. Some statistical methods, Bayesian estimation method, information maximizing strategy and the multi-armed bandit theory, are introduced according to the RoboCup regulation. Finally, an observer agent is designed to implement these features. This program is written in JAVA with multi-threading to realize both reactive response and deliberative planning.

Introduction

In order to study cooperation of multiple agents, soccer was proposed as a benchmark problem(Kitano *et al.* 1995). One of the most important research issues is how to realize robust and good performance of multi-agent systems in dynamic situations. It requires reactive and quick response along with deliberative planning during the game. Although reactive performance is effective in getting points in the game, its effectiveness also depends on the opponent's strategy. In order to keep the advantage against any opponent team, the adaptability to opponent players is important. This capability is considered a key to robustness for intelligent behavior in tournament competition.

Identification of an opponents strategy is a significant research issue for this purpose. An opponents model is formalized statistically so that any kind of player can be classified with an internal representation. An opponent is characterized as the conditional probability of selecting an action in a particular situation. Therefore, identification of an opponent's strategy is studied as a problem of estimating a probability distribution. Moreover, a player's perception capability under the RoboCup regulation is restricted naturally like real-world environment. Complete observation is not guaranteed in this framework. Thus, they should not only get points, but also make appropriate actions to get sufficient information to predict an opponent's strategy. This problem is called active perception.

In this paper, some statistical techniques to solve these problems are proposed. Firstly, Bayesian estimation method is formalized in the RoboCup environment. Secondly, an active perception based on information maximizing strategy, which derives information from the environment as much as possible, is discussed. Thirdly, a theoretical approach called the multi-armed bandit problem is introduced into the RoboCup situation. It is also concerned with the trade-off between observation and getting points. Finally, an observer agent to identify an opponent's strategy in the game is designed. It is implemented in JAVA using a multi-thread mechanism. It is also a key technique to realize both reactive action and deliberative action in a agent program.

Bayesian statistical formalization

Strategies of multi-agent system are constructed with the following three factors:

1. each player's strategy(behavior in each situation);
2. architecture of the team; and ,
3. process flow among team mates(carrying ball to a goal).

As a first step of analyzing opponent's strategy, identification of each player's strategy is the focus here, and it is formalized as a Bayesian statistical estimation problem.

Opponent's strategy

Most of soccer client programs are constructed from a combination of rules such as:

IF ball-distance $< l_a$ *and goal-distance*$< l_b$,
THEN kick to goal-direction by power P.

The precondition part: ball-distance $< l_a$ and goal-distance $< l_b$, is considered as a particular situation. In this case, it is represented with specific parameter l_a, l_b which can't be known by the observer agent. The postcondition part: kick to goal-direction by power P, is considered as a particular action of the player. In general form, player agents can be denoted with situation set $S = \{s_1, \cdots, s_m\}$, action set $A = \{a_1, \cdots, a_n\}$

and rule set which is any combination of statements of the form:

$$rule: IF\ s_i\ THEN\ a_j.$$

However, it is significantly difficult to identify the opponent's A, S and rule sets exactly from observation of the game. The observer agent has its own internal representation of situation $s_i \in S = \{s_1, \cdots, s_m\}$, and action $a_j \in A = \{a_1, \cdots, a_n\}$, but they might not agree with opponent's $S$and A. So, an approximating method of the opponent's model is necessary.

In this case, the approximated model can be described as a non-determinant model. When a situation is recognized as s_i, it can be considered that an action a_j arises from an unknown player. It is observed as a statistical event, and it can be recorded as the frequency of the action in the situation s_i. If this frequency is asymptotically stable, we can consider asymptotic frequency as the probability of the action a_i under the situation s_i, $P(a_j|s_i)$. Thus, we can understand that the action a_j of this player is selected by a conditional probability $P(a_j|s_i)$. Any opponent player's strategy which is a selection mechanism of each action in any situation can be characterized as a particular conditional probability distribution $P(A|S)$. So, identification of an opponent strategy is considered through a estimation of conditional probability distribution $P(A|S)$ with observer's internal representation (A,S).

Bayesian formalization is introduced for estimation of probability distribution. Let's describe sensing data at time t as d^t, and a set of $\{d^1, \cdots, d^t\}$ as D^t. The conditional probability of action a_j in the situation s_i which is result after d^t comes is denoted as $P(a_j|D^t, s_i)$. At first, $P(A|D^0, s_i)$ is set to uniform distribution as non-informative prior. (For example, all $P(a_j|D^0, s_i) = \frac{1}{n},\ ,(j = 1 \cdots, n)$). When data d^t comes, it is updated as the posterior probability,

$$P(a_j|D^t, s_i) = \frac{P(d^t|s_i, a_j)P(a_j|D^{t-1}, s_i)}{\int_A P(d^t|s_i, a_j)P(a_j|D^{t-1}, s_i)da_j}. \quad (1)$$

Here, $P(d^t|s_i, a_j)$ is a likelihood, which is the degree of plausibility that data d^t comes from action a_j in the situation s_i. It can be calculated by the observer agent in the latter section.

There are two reasons to take Bayesian formalization. Firstly, if we consider probabilistic modeling in a non Bayesian way(frequentist's way), $P(A|S)$ should be computed from only the frequency of a player's action. However, in the small sample area, frequency is too rough to evaluate opponent's strategy. For example, if an action can't be observed in the situation, the probability of the action is equal to zero. Secondly, in Bayesian way, the action in the same case is estimated as $\frac{1}{n}$. So, Bayesian formulation is more natural than frequentist's formulation in this case. Secondly, even in the large sample case, the event that an opponent performs action a_j in the situation s_i is hard to be observed, because received data can't be mapped on to

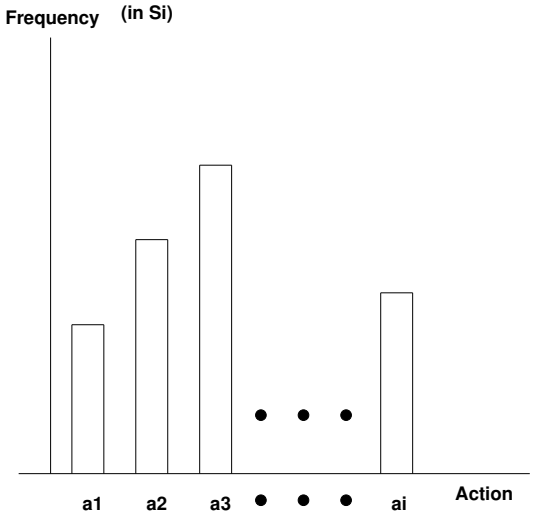

Figure 1: Asymptotic frequency of action

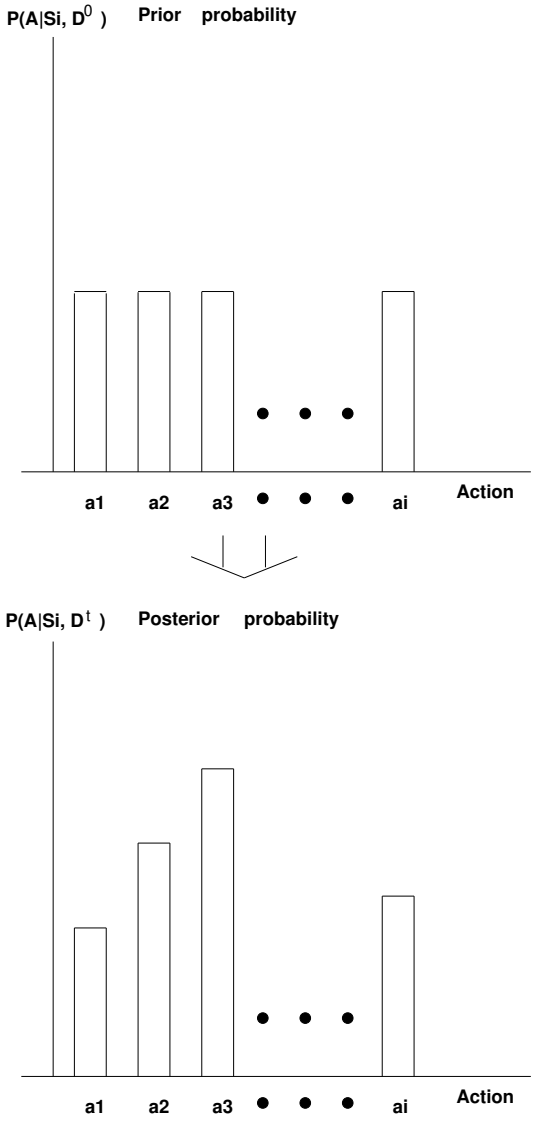

Figure 2: Prior and Posterior probability of action

a particular action directly. The raw data is evaluated through likelihood of the action. So, Bayesian way, which computes posterior probability with likelihood and prior probability, is necessary.

Active perception

A player can watch received data sent from soccer server. However, it is not enough to estimate the other player's strategy in one time slice as observation time is very limited during the game. So, the observer agent should collect sufficient evidences effectively. According to this line, active perception has been studied in a lot of work based on the statistical decision theory, for example, the active vision system which makes decision where to look next on a dining table(Rimey & Brown 1994). In this work, active perception is done in order to optimize information gain. In the case of the soccer game, observer agent should collect information efficiency so as to identify opponents as quickly as possible. So, a criterion to estimate information gain of data is proposed. The entropy of a player p in situation s_i is defined as:

$$H^t_{p,s_i} = -\sum_{j=1}^{n} P(a_j|D^t, s_i) \log P(a_j|D^t, s_i). \quad (2)$$

It indicates the degree of uncertainty of a player's action in the situation s_i. Before no data comes, the entropy of non-informative prior takes the maximum value, and the entropy is decreased by each coming data. So, the difference of entropy as the following is considered as the information gain of d^t.

$$I^t_{p,s_i} = H^{t-1}_{p,s_i} - H^t_{p,s_i} \quad (3)$$

These entropy and information gains are used as criteria to select an action. Active perception schemes are described as follows:

1. Following the particular opponent who has the largest total entropy of all opponents,
2. Choose the particular situation which has the largest entropy of all situations.

It can be called information maximizing strategy. These criteria can be also used to judge whether classification of S and A is proper or not. This feature is helpful in what is called *emergent approach.*

Multi-armed bandit theory

Observation-Optimization trade-off

The ultimate purpose of soccer game is to get more points than the opponents, so we have to choose the criterion which reflects the purpose of winning the game. On the other hand, in order to identify the opponents' strategy, we should collect as much information about opponents as possible. However, these two objectives sometimes contradict. For example, if it takes a lot of time to perceive the opponents behaviors, we might lose the game.

Such kinds of problems appear in game theory. For instance, bandit theory(Berry & Fristedt 1985) aims to get as much rewards as possible, with some experiments to get information about the multi-armed slot machine.

The bandit theory, in the simplest case, deals with the arms of a two-armed slot machine. At each "pull of an arm", the user receives a reward chosen by a random mechanism which is different for each arm.

The traditional bandit theory needs a lot of assumptions. For example, the probability distribution of the reward belongs to a certain parametric family, or the reward sequence is assumed to be independently identically distributed (i.i.d.). Since such assumptions are not practical in complicated situations like the soccer game, we apply the nonparametric bandit theory(Lai & Yakowitz 1995) described below which only needs few weak assumptions.

Nonparametric bandit theory

We have k choices of actions at each time step t. When we choose the j-th action, a reward $X^{(j)}_t$ is provided. We assume that a global average for each choices exists,

$$\lim_{n\to\infty} \frac{1}{n} \mathrm{E}[X^{(j)}_1 + \cdots + X^{(j)}_n] = \mu_j, \quad (4)$$

In the present framework, we do not assume that the distribution of $X^{(j)}_t$ belongs to some parametric distribution nor $(X^{(j)}_1, X^{(j)}_2, \ldots)$ is i.i.d. sequence. So we can apply the theory to nonstationary and time-dependent sequences.

Under several additional regularity conditions, the regret

$$R_n = \sum_{j:\mu_j<\mu^*} (\mu^* - \mu_j)\, \mathrm{E}[T^{(j)}_n] \quad (5)$$

is of the order $O(\log n)$ as the number of trials $n \to \infty$, where $T^{(j)}_n$ is the number of cases that the j-th action is selected until n and $\mu^* = \max_j \mu_j$; hence, R_n is the expected difference between the reward obtained by optimal sequence of actions with complete information and the reward obtained by the algorithm.

The adaptive opposites

If the opposites are more intelligent, they will change their strategy according to the player's action. The 'lob-pass' problem(Abe & Takeuchi 1993) that is an extension of the bandit problem deals with such adaptive opposites explicitly. The original idea of the 'lob-pass' problem is motivated by playing tennis. The player has two choices of actions, 'lob' and 'pass'. If the player plays lob so often, the success rate of lob decreases and the success rate of pass increases. On the other hand, if the player plays passe so often, the success rate of pass decreases and the success rate of lob decreases.

This framework will also work for the soccer game. In the present formulation, those success rates only depend on the frequency of the past actions. Namely, the success rate of pass is given by

$$\text{Prob(lob success)} = a_1 r + b_1, \quad (6)$$

$$\text{Prob(pass success)} = a_2 r + b_2, \quad (7)$$

where r is the frequency of 'lob' and $a_1 < 0, a_2 > 0$ are unknown parameters.

Abe(Abe & Takeuchi 1993) analyzed the regret of the lob-pass problem under several kinds of assumptions. Currently, the theory guarantee the upper bound of the regret only for the case that the number of choices is two and the binary reward (success or fail) is given. It also needs several assumptions about the model.

Situation dependent case

In the framework described above, we fixed the situation where the player chooses an action. But the player will face a lot of different situations in a actual game, and the distribution of the reward also changes.

One simple idea to deal with the situation dependency is that the player learns environment and plays actions independently for each situation, namely, the action of the situation α is decided from the history of the situation α,

$$\frac{1}{T_n^{(j)}(\alpha)}\{X_{t_1}^{(j)}(\alpha) + \cdots + X_{t_{T_n(j)}(\alpha)}^{(j)}(\alpha)\}. \quad (8)$$

where t_i is the i-th step in which the action j is selected in the situation α.

But this idea will work only for the case of a few number of situations. If two situations are similar, the distributions of reward for actions will be also similar. Therefore, suppose each situation is represented as a point in a distance space, and the player learns and plays in several points. If the player faces a novelty situation, the player choose an action according to the linearly interpolated history. Namely, suppose the player has already experienced the situations $(\boldsymbol{s}_{t_1}, \ldots, \boldsymbol{s}_{t_{T_n^{(j)}}})$ for the action j and got rewards $(X_{t_1}^{(j)}, \ldots, X_{t_{T_n^{(j)}}}^{(j)})$. From those observations, we obtain the linear regression function,

$$\hat{X}^{(j)}(\boldsymbol{s}) = \boldsymbol{a}^{(j)\mathrm{T}}\boldsymbol{s} + b^{(j)}, \quad (9)$$

where $\boldsymbol{a}^{(j)}$ and $b^{(j)}$ denote regression parameters. When the player is in a new situation $\boldsymbol{s}_t$, the player can decide the action by using the interpolated reward $\hat{X}^{(j)}(\boldsymbol{s}_t)$ instead of (8).

However, the regret is not guaranteed to be of the order $O(\log n)$ anymore. As a future work, we should integrate the situation dependency and the adaptive opposites. The program for multi-armed bandit theory will be available at the URL, http://www.etl.go.jp/etl/suri/akaho/Bandit.html .

Current buffer				Previous buffer			
ball	t_a	*dist*	*dir*	ball	t_b	*dist*	*dir*
goal l	t_c	*dist*	*dir*	goal l	t_d	*dist*	*dir*
goal r	t_e	*dist*	*dir*	goal r	t_f	*dist*	*dir*
opp 1	t_g	*dist*	*dir*	opp 1	t_h	*dist*	*dir*
opp 2	t_i	*dist*	*dir*	opp 2	t_j	*dist*	*dir*
:	:	:	:	:	:	:	:

Table 1: Double buffered memory

Observer Agent

According to the above discussion, the observer agent is designed. It communicates with the RoboCup soccer server like a usual soccer client (player), and it observes and estimates the opponent player while playing the game.

Internal representation

Visual data coming from the *Soccer server*(Noda 1996) involves each object data shown at time t with it's direction and distance from the observer's point of view:

(see t ((*ball*) *dist dir vdir*) ((*goal*$_r$) *dist dir vdir*) ...).

Visual data which is segmented sequentially into unit of time is not enough and the observer's field of view is limited. Moreover, movement information for each the object can't be obtained from raw data. So, double buffered memory for every object is proposed in order to collect data about every object shown in the past. One buffer keeps the newest data and the other buffer keeps previous data to the newest one(Figure 1). When the observer sees more than two land marks(fixed objects such as a goal, flag, etc.), absolute coordinates (x, y) are also obtained by simple geometric calculation. Movement information(direction and velocity) of an each object is derived from two data of different time in the double buffered memory. Thus, an observer can also make use of this information as d^t which is calculated from raw data at time t.

In order to realize situations for the opponent player, visual data d^t from observer's point of view should be translated to visual data from the opponent's one. Visual data is put on to a relative two dimensional space which takes the observer's position as the origin of the coordinate. If the origin moves to any other player's position, it can be considered as the visual data viewed from the other player's position.

In current stage of this study, Situation set S is decided at first as extract information of d^t. Action set A is also given at first. However, this classification of situation and action problem is essential for the research(Noda, Matsubara, & Hiraki 1996). As future work, it is planned that S and A is generate through the game in RoboCup environment.

The likelihood of d^t, $P(d^t|a, s)$ is computed by a specific function that depends on d^t, a, and s. This function is decided experimentally in the RoboCup situation because it depends on the soccer server.

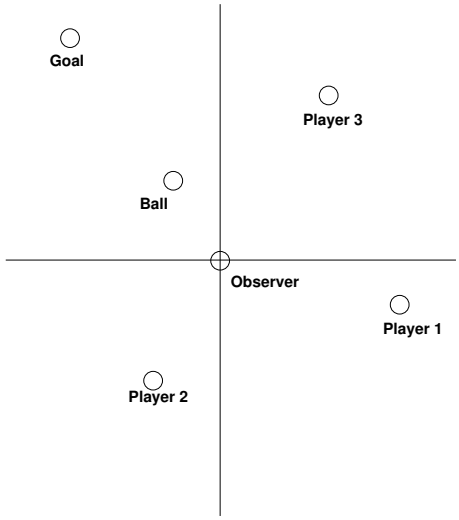

Figure 3: position data from observer's point of view

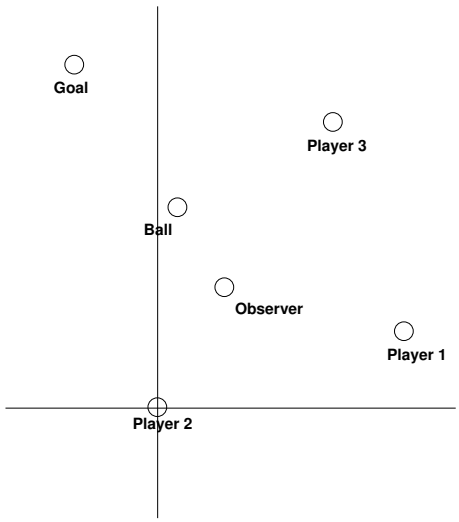

Figure 4: position data from player2's point of view

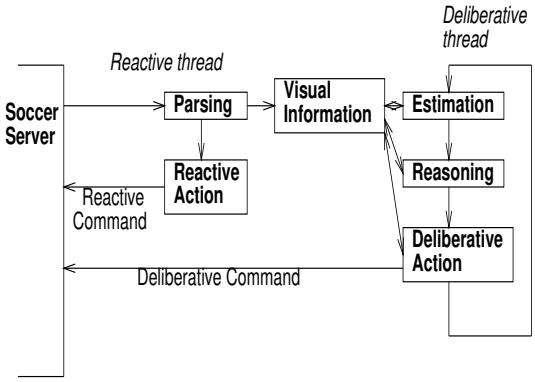

Figure 5: Multi-thread in Observer Agent

	a_1	a_2	a_3	$\cdots$	a_m
s_1	$p(a_1\|s_1)$	$p(a_2\|s_1)$	$p(a_3\|s_1)$	$\cdots$	$p(a_m\|s_1)$
s_2	$p(a_1\|s_2)$	$p(a_2\|s_2)$	$p(a_3\|s_2)$	$\cdots$	$p(a_m\|s_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$

Table 2: Conditional probability distribution $P(A|S)$

Conditional probability which is updated by equation (1) is represented in Table 2, and also it is interpreted as an estimation of the opponent's strategy.

Multi-threading

The task of estimating the opponent's strategy should be executed deliberatively. On the other hands, behavior of active perception should be executed reactively to the data input cycle of soccer server. These different tasks are implemented by multi-thread technique in JAVA language (Figure 5). This program and related information will be available at the URL, http://www.etl.go.jp/etl/suri/motomura/soccer.html .

Conclusion

In this paper, Bayesian statistical formulation has been introduced in the context of RoboCup. By this framework, entropy and information is defined naturally, so information maximizing strategy for active perception can be used. Moreover the multi-armed bandit theory has been introduced. Finally, an implementation has been shown for the observer agent, which identifies the opponent's strategy.

Acknowledgements

The authors thank to Dr. Steven Phillips, who gives many helpful comments, and Dr. Itsuki Noda, who provides source codes of the soccer server and clients.

References

Abe, N., and Takeuchi, J. 1993. The 'lob-pass' problem and an on-line learning model of rational choice. In *Proc. of COLT'93.*

Berry, D., and Fristedt, B. 1985. *Bandit Problem.* Chapman and Hall.

Kitano, H.; Asada, M.; Kuniyoshi, Y.; Noda, I.; and Osawa, E. 1995. Robotcup: the robot world cup initiative. In *IJCAI-95 workshop on Entertainment and AI/Alife.*

Lai, T., and Yakowitz, S. 1995. Machine learning and nonparametric bandit theory. *IEEE trans. on Automatic Control, vol40, No.7.*

Noda, I.; Matsubara, H.; and Hiraki, K. 1996. Learning cooperative behavior in multi-agent environment. In *Proceedings of 4th Pacific Rim International Conference on Artificial Intelligence.*

Noda, I. 1996. http://www.ci.etl.go.jp/ñoda/soccer/.

Rimey, R., and Brown, C. 1994. Control of selective perception using bayes nets and decision theory. *International Journal of Computer Vision 12:2/3.*