The GPR-2D 2013 Team Description Paper: Reinforcement Learning to Improve Attack decision-making and attack strategy modification during a match.

João Alberto Fabro, André Luiz Constantino Botta,
Gyovann Albert Pinto Parra
Computational Intelligence Research Group, Informatics Department (DAINF)
Federal University of Technology - Paraná (UTFPR)
Curitiba, PR, Brazil
e-mail: fabro@utfpr.edu.br, {alcbotta, gyo.albert}@gmail.com

José Rodrigo Ferreira Neri
Department of Automation and Systems (DAS)
Federal University of Santa Catarina (UFSC)
Florianópolis, SC, Brazil
e-mail: jrneri@das.ufsc.br

***Abstract* — GPR-2D, that stands for "Grupo de Pesquisa em Robótica" (Robotics Research Group as written in portuguese) is a Robocup 2D simulation team that uses the Q-Learning technique to devise a better control action selection when close to the adversary goal. This document presents the proposed reinforcement learning approach, and the advances since 2012. The original proposal used a continuous learning procedure, where the GPR2D team tried to adapt itself to each opponent team during the match. The main advance for this year is an alternating approach, where previous trained solutions, that worked well against different teams, are tried during a difficult match, aiming to bring previous successful experiences to each match. The approach is similar to a coach that changes the tactics during a match, in order to surprise the adversary. Simulations results show that the proposed approach retains the good characteristics of the previous one but also can uses previously learned behavior to find new solutions if it cannot score goals.**

Keywords - Simulated robot soccer, machine learning, Q-Learning algorithm.

1 - Introduction

GPR-2D is the Robotic Research Group ("***G**rupo de **P**esquisa em **R**obótica*" as written in portuguese) robotic soccer team. Its development started in 2007, and its first competition was the national league in 2009. Already in 2011, it reached the championship in Brazil, and last year it was ranked 3rd at the Latin American Robotics Competition - LARC [1], and participated at the Robocup, achieving the 9th~12th position. The aim of our initiative is to research methods and algorithms that can be applied to any group of cooperative robots, and the results obtained within the 2D simulation category are being transposed to the 3D simulation category team (that obtained the 4th position at LARC last year), and also to the Small Size category team

, p. 1, 2013.

with real robots. Several other real robots that are in development at the university can also benefit from the algorithms and techniques under development. The Robocup 2D simulation category focuses on the development of distributed control algorithms, so each of the simulated agents should be able to represent the environment, select sub-objectives, plan and implement the plans through a decision-making procedure. As these objectives depend upon individual decisions taken by each player, the approach proposed by the GPR-2D team was to use Reinforcement Learning (Q-learning [2]) to modulate the decision process of each simulated player, based on training during the matches. This algorithm was proposed and used in the GPR-2D team that achieved the first place of the Brazilian RoboCup 2D simulation category in 2011[3], and third place in the Latin American Robotics Competition in 2012 [1]. The main advances difference that was developed in this last year is the use of different behaviors (different Q-learning matrices) during a single game. Against different teams, different approaches lead to more goals scored. But the previous approach used only one set of matrices for the Q-learning, that was updated and saved at the end of each game, and therefore used on the next game. The main advance in this year's approach is the use of various matrices, trained against different teams previously, that are tried during a single match, if the results aren't favorable. This paper briefly presents the original Q-learning approach used by the team for 2012 Robocup (section 2), and presents the proposed variations in section 3. Section 4 presents some simulation results, and section 5 draw some conclusions and discuss the results.

2 - The GPR-2D Team

The GPR-2D team started in 2007 at the State University of West-Paraná (UNIOESTE). In 2009, it first participated at the Brazilian national competition, in collaboration with the Federal University of Technology - Paraná (UTFPR). Nowadays, the team is a joint effort of both UTFPR and Federal University of Santa Catarina (UFSC) (all from Brazil). The team is based on the Agent2D source code, provided by the Helios team (Japan) [4]. The difference resides in the decision-making process that occurs once the team has reached the attack (i.e., the player that is in possession of the ball is inside the goal area of the adversary team – see fig. 1). Q-learning [2], a reinforcement learning algorithm, is used to modulate the agent's behavior when inside one of these areas. This strategy allows the simulated agents to select the best action when in possession of the ball, and inside one of the areas of fig.1. The reinforcement takes place each time the team scores a goal, and sequences of actions leading to a goal are reinforced. Details of the learning procedure are presented in the following sub-sections, and a more thorough explanation can be found in [5].

2.1 Reinforcement Learning

Reinforcement Learning (RL) is a technique that allows the learning of decision-making policies. In the case of robotic soccer, each agent has a variety of options regarding its next action: to move (with or without possession of the ball), to execute a pass to a fellow player, or to kick the ball towards the adversary goal.

Q-Learning

The Q-learning algorithm [2] is used to find an optimal policy for a given situation interactively, using a state-action matrix. The matrix identifies what the agent should do (action) when in each state. At first, all valid actions have the same probability of being chosen at each state. By providing a reward for some actions that are better for the agent in some sense, the matrix slowly converges for the best solution, that optimizes the decisions for the agent. The matrix is represented by a function Q that is learned by agent, and after learning, the agent knows which action gives the greatest reward at each specific state. The function Q(s,a) of expected reward is learned through successive execution of the reinforcement algorithm, using the following equation:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \gamma \max_a Q(s_{t+1}, a_t) - Q(s_t, a_t)] \quad (1)$$

where s_t corresponds to the current state, a_t is the action taken at state s_t, α is the learning rate, r_t is the reward received by taking the action a_t at the state s_t, γ is the discount factor and $max_a\ Q(s_{t+1}, a_t)$ is the utility of state "s" resulting from taking action "a". In the experiments, α was set to 0.5 and γ to 0.8. The function $Q(s_t, a_t)$ is the value associated with the state-action pair (s_t, a_t) and represents how good is the choice of this action in maximizing the cumulative return function. The action-value function $Q(s_t, a_t)$, that store the reinforcements received, is updated from its current value for each state-action pair.

2.2 Problem Modeling

To apply the Q-Learning algorithm, the first step is the discretization of the states that each simulated player can achieve. The proposed states for this problem were:

Lead_pass	In this state, it is possible to launch the ball to a fellow player better positioned;
AdversaryBehind	The closest adversary is behind the player that is in possession of the ball
AdversaryFarAway	The distance to the closest adversary is bigger than 7 meters
AdversaryNotSoClose	The distance to the closest adversary is bigger than 6 meters
AdversaryIsClose	The distance to the closest adversary is bigger than 5 meters
AdversaryVeryClose	The distance to the closest adversary is bigger than 4 meters
KickOpportunity	There is a direct line from the player and the adversary goal

The following are the possible actions that can be taken by each agent:

Launch	Pass the ball towards a fellow player that is much closer to the adversary goal
SlowForward	To slowly advance with the ball
FastForward	To quickly advance with the ball
Pass	Pass the ball to a fellow player that is close, but not necessarily best positioned
Dribble	Try to dribble the closest adversary player
Hold_the_ball	Just hold the ball
Kick	To kick the ball in the direction of the adversary goal

Thus, seven states were defined for the environment, and seven actions for the agents. After defining the states of the environment and actions of the players, it is necessary to define the matrices R (Reinforcement Matrix) and Q. The matrix R is where the reinforcements are stored, in order to define the actions of the agent. In the Q matrix the values of $Q(s_t, a_t)$ assigned by Q-learning algorithm are stored.

Two interest areas were defined on the field, as can be seen on Fig. 1. For each of these areas, it were defined a set of Q and R matrices. The main difference between this approach and the base Helios team is the use of the Q-learning matrices to decide the actions when the agents enter the interest areas, thus providing the ability to "learn" how to score goal against different adversaries.

Fig. 1. The two different interest areas (each one has different Q-Learning Matrices)

Only when the players of our team had the possession of the ball, and approached the adversary goal (thus entering areas 2 or 1), the matrices were updated. After these first 10 matches, the stabilization of the behavior of the attack was perceived.

3 – Proposed Advances for 2013

The main advance proposed for this year's competition is the use of multiple reinforcement matrices during the same match. One of the drawbacks of the reinforcement learning approach is that the reinforcement only occurs, in our team, when a goal is scored. If the adversary team is successful in avoiding this, no reinforcement is made to the Q-learning matrices during the match. For that reason, we started to reset the learning matrices at the begin of each game, to provide a wider range of possibilities, since every action would have the same probability of happening at each new game. But since the reinforcement learning is a procedure with slow convergence, usually a single game with its 6000 cycles wasn't enough to provide a meaningful learning process. We perceived that the convergence of the learning matrices took place after 10 games against the same team, in average. So we decided to store the learning matrices obtained against selected opponent teams after

10 consecutive games, and alternate the use of these matrices during a match, if our team is not scoring goals. The strategy used is simple: the team starts with a certain set of matrices and, if it is not capable of scoring goals after a certain interval, a new set of matrices, result of a training against a different opponent team, is loaded. This strategy allows to use behaviors that are successful against different opponents in a single match. If the policy in use is successful in scoring goals, the matrices are reinforced, and there is no alternation of strategy. In the next section, some results are presented that shows the results obtained by the proposed approach.

4 - Simulations Results

The initial training of the policy was executed during 10 matches against two teams: the base team (agent2d version 3.1.1) [4] and ITAndroids [6]. These teams were used for the training phase because it was common the scoring of goals, providing opportunities for the Q-learning to execute and reinforce the best actions. For each of these teams, the training phase consisted of the sequential execution of 10 matches against the same team, continuously updating the learning matrices using the Q-Learning algorithm. At the end of this phase, two sets of matrices were obtained. To evaluate the performance of the proposed approach, 10 simulations were performed against the teams presented on Table I, changing the matrices at the beginning of the second half if GPR was unable to score during the first half. The results are shown on Table I.

Table I - Results of the Proposed Approach

Adversary Team	Matches won by GPR (of 10)	Sum of scores (GPR x other team)	Comments
Helios_Base (agent2d_3.1.1)	8	21 x 8	Only lost 2 games, both 0x1
ITAndroids 2012	6	11x12	Champion of Latin America in 2012.
FCPortugal 2012	5	21x18	9th at Robocup 2012
Gliders 2012	3	11x20	4th at Robocup 2012
MarIik 2012	1	3x16	3rd at Robocup 2012
WrightEagle 2012	0	1x39	2nd at Robocup 2012
Helios 2012	0	1x39	1st at Robocup 2012

It can be seen from the results that the proposed approach outmatches Agent2D, the team used as base for the development, and achieves a performance that compares the team to ITAndroids [6] (Latin American Champion in 2012) and FCPortugal [7], that achieved the 9th place at Robocup 2012. The team isn't comparable yet to the 4 best placed teams of 2012.

5 – Conclusions and Discussion

This paper presented the approach used by the GPR-2D team of simulated soccer agents. The agents use the Q-learning algorithm to select actions for the player with

ball possession inside the adversary goal area, adapting the behavior in order to achieve better results, scoring more goals. The agents continuously learning process allows the team to adapt during the matches, searching for strategies that lead to situations where more goals are scored. But as the learning process is relatively slow, if the team faces a very strong opponent, and cannot score goals at all, there is no learning, and the team does not adapt. In this situation, a change of tactics is needed. In this years team, if the team is unable to score a goal after 3000 cycles of simulation, a different strategy is implemented, that can give renewed chances for the team to score and win difficult games. The different strategies are represented by different Q-learning matrices, trained against diverse adversaries, and provide opportunities for improved performance of the team.

From the presented results it is clear that the team isn't capable to compete with the best teams from the Robocup, since it was unable to win matches against the 3 top teams from last years competition, and even score a single goal against the top 2. This is due to a characteristic of the approach presented: it improves the ability of the team to score goal when inside the adversary goal area. Both WrightEagle[8] and Helios[4] have strong defensive approaches that do not allow the ball to enter the goal areas, so the Q-learning matrices rarely have any opportunity to be executed.

Future improvements of the approach include the use of other reinforcements, for instance when a pass is successful, and the expansion of the area of interest for the entire forward field. The penalty routine should also be improved. It is also planned to alter the positioning of the team in some situations (such as when the team is losing a game) to provide more opportunities to score goals.

References

1. 2012 Latin American Robotics Competition (LARC) Results Page, online, available at: http://www.cbrobotica.org/Results.htm, Accessed: January 2013.
2. Dayan, P. Technical Note Q-learning, Centre for Cognitive Science, University of Edinburgh, Scotland: University of Edinburgh, (1992).
3. Brazilian Robotics Competition Results Page, online, available at: http://www.cbr2011.org/Sim2D.htm, Accessed: March 2012.
4. Hidehisa Akiyama, Helios RoboCup Simulation League Team, online, available at: http://rctools.sourceforge.jp/pukiwiki/, Acessed: January 2013.
5. Neri, J.R.F.; Zatelli, M.R.; Farias dos Santos, C.H.; Fabro, J.A.; , A Proposal of QLearning to Control the Attack of a 2D Robot Soccer Simulation Team, 2012 Brazilian Robotics Symposium and Latin American Robotics Symposium (SBR-LARS), pp.174-178, 16-19 Oct. 2012.
6. Mello, F.; Ramos, L.; Maximo, M.; Ferreira, R. and Moura, V. ITAndroids 2D Soccer Simulation Team Description Paper, online, available at: http://www.socsim.robocup.org/files/2D/tdp/RoboCup2012/, Accessed: January 2013.
7. Lau, N.; Reis, L. P.; Mota, L.; Almeida, F. FC Portugal 2D Simulation: Team Description Paper, online, available at: http://www.socsim.robocup.org/files/2D/tdp/RoboCup2012/, Accessed: January 2013.
8. Bai, A; Zhang, H.; Lu, G. ; Jiang, M. and Chen, X. WrightEagle 2D Soccer Simulation Team Description 2012, online, available at: http://www.socsim.robocup.org/files/2D/tdp/RoboCup2012, Accessed: January 2013.