Opponent Modeling against Non-stationary Strategies (Doctoral Consortium)

Pablo Hernandez-Leal
Supervisor: Enrique Munoz de Cote and L. Enrique Sucar
Instituto Nacional de Astrofísica, Óptica y Electrónica
Sta. Marian Tonantzintla, Puebla, México
pablohl@ccc.inaoep.mx

ABSTRACT

Most state of the art learning algorithms do not fare well with agents (computer or humans) that change their behaviour in time. This is the case because they usually do not model the other agents' behaviour and instead make some assumptions that for real scenarios are too restrictive. Furthermore, considering that many applications demand different types of agents to work together this should be an important problem to solve. We contribute to the state of the art with opponent modeling algorithms. In particular we proposed 3 approaches for learning against non-stationary opponents in repeated games. Experimentally we tested our approaches on three domains including a real world scenario which consists of bidding in energy markets.

Keywords

Non-stationary strategies; opponent modelling; Markov decision process; learning

1. INTRODUCTION

Current learning techniques do not fare well with agents that change their behavior during a repeated interaction. There is one thing in common in cooperative or competitive scenarios: agents must learn how their counterpart is acting and react quickly to changes in their behaviour. The proposed research is focused on learning non-stationary strategies, given that agents may use different strategies and switch among them. Dealing with non-stationary opponents involves three different aspects: (i) Learning a model of the opponent. (ii) computing a policy (plan to act) against the opponent (since the objective is to maximize the rewards throughout the interaction) and (iii) detecting switches in the opponent strategy. Notice that under our proposed setting, an already optimal policy will be suboptimal if the opponent changes its strategy.

1.1 Motivation and justification

There are diverse reasons why to study non-stationary opponents in multiagent systems, for example: (i) Predicting the behavior of other agents is crucial in competitive environments since opponent models can be used to identify weaknesses. (ii) On the other side, learning models of agents in cooperative environments can be useful to perform optimal planning of a collaborative task. (iii) Human-agent interaction is a current research area that designs algorithms for humans interacting with computer agents. The bottom line is, it does not matter whether the agent and the human want to work together or compete against each other, agents often change behaviours through time.

1.2 Research questions

In order to pursue our thesis we formulate the following questions: (i) How should we model non-stationary agents? (ii) What is an efficient way to detect strategy switches in the opponent? (iii) What should be taken into account when designing a planning algorithm for long-term interactions when a model of the opponent/teammate is at hand?

1.3 Problem setting

One learning agent $\mathcal{A}$, and one opponent $\mathcal{O}$ share the environment. Both agents take one action (simultaneously) in a sequence of rounds. They both obtain a reward r that depends on the actions of both agents. The objective of agent $\mathcal{A}$ is to maximize its cumulative rewards over the entire interaction. Agent $\mathcal{O}$ has a set of M possible strategies to choose from and can switch from one to another in any round of the interaction. A *strategy* defines a probability distribution for taking an action given a history of interactions. We have experimented in well known games as the iterated prisoner's dilemma (iPD) and bilateral bargaining.

2. STATE OF THE ART

Current approaches have different limitations. Game theory has developed several algorithms in this area, their problem is that they focus on finding Nash equilibrium implying rational (perfect) agents on all situations. Moreover, they are not designed for non-stationary opponents. Behavioral game theory put special focus on bounded rational agents, however they mostly use single-shot games to derive its models and experiments. Decision theoretic planning algorithms assume there is a single agent in the environment. In reinforcement learning algorithms the step from a single agent to multiagent is not straightforward and current approaches require a large number of interactions to learn efficiently.

3. CONTRIBUTIONS

Our first contribution is the MDP-CL framework [3] designed for learning and planning against non-stationary opponents in repeated games. A Markov decision process is

Appears in: *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015), Bordini, Elkind, Weiss, Yolum (eds.), May 4–8, 2015, Istanbul, Turkey.*
Copyright © 2015, International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

learned as a model of the opponent and solving it gives us the optimal policy against it. For detecting switches, models are learned in different windows of interactions and comparisons among them reveal when models are changing.

Then, we take into consideration what happens when a priori information can be obtained when facing non-stationary opponents. We propose two extensions for MDP-CL [4]: (i) A priori MDP-CL uses prior information (in the form of a set of opponent models) to quickly detect the opponent model. (ii) Incremental MDP-CL learns new models from history of interactions and it will not discard them once it detects a switch. In this way it keeps a record in case the opponent reuses a previous strategy, thus reducing the need of relearning the model.

From our results we noticed that one of the important aspects needed for an algorithm to be successful against non-stationary strategies is the exploration process. Thus, a successful learning algorithm should not only detect switches consistently; but should also explore the state space efficiently and encourage revisiting far-visited state-action pairs. This is a different type of exploration designed for switch detection which we coined as *drift* exploration. Our second contribution is R-MAX# (read R-MAX *sharp*, since it is sharp to changes), a novel algorithm based on R-MAX [1] that updates the model through the complete interaction performing an implicit drift exploration, in order to be used against non-stationary opponents.

Finally our third contribution is DriftER [5] an approach based on concept drift that detect changes in non-stationary opponents in an efficient and practical way. DriftER performs a monitoring of the quality of the learned model and use that as a indicator for switch detection.

4. EXPERIMENTS AND RESULTS

We used three domains for performing experiments: the iterated prisoner's dilemma, a negotiation task and periodic double auctions inside the PowerTAC simulator [6].

4.1 Switch detection

MDP-CL was compared against an offline reinforcement learning technique for non-stationary environments (HM-MDPs) [2] in the iPD game. Some conclusions from the experiments are: MDP-CL is an on-line learning approach, which do not need to know before hand the number of possible strategies of the opponent. Also it computes the policy in a faster way since solving a MDP is computationally much simpler than solving the HM-MDP.

4.2 Drift exploration

We tested two different domains in which drift exploration is necessary to obtain an optimal policy —due to the non-stationary nature of the opponent's strategy. The iPD and a negotiation task. In both scenarios, the use of switch detection mechanisms were not enough to deal with switching opponents, since they do not perform any drift exploration. Our approach, R-MAX#, which implicitly handles drift exploration is generally better equipped to handle non-stationary opponents of different sorts. Its pitfall lies in its parameterization (the parameter which controls when to re-explore the space state), which generally should be large enough so as to learn a correct opponent model, yet small enough to react promptly to strategy switches.

4.3 PowerTAC

Previous experiments were performed on two domains. However, to continue increasing the complexity of the domain we performed experiments in a more realistic scenario such as double auctions in energy markets. In this context, PowerTAC can be used to perform research on retail energy markets. The previous champion of the competition was not capable of adapting quickly to non-stationary opponents (which change from one stationary strategy to another), impacting their total profits. In contrast, DriftER obtained better scores in terms of profit and accuracy than the previous champion of the competition against switching opponents.

5. CONCLUSIONS

When two agents interact for long time they probably change strategies. This renders the problem non-stationary in which most algorithms fail. We contribute to the state of the art with (i) a learning algorithm (MDP-CL) designed for non-stationary opponents in repeated games. (ii) An algorithm (R-MAX#) which performs an efficient re-exploration of the space state for detecting switches in the opponent. (iii) DriftER approach which is based on concept drift ideas to detect switches. We performed experiments in two simple domains and one realistic domain: the PowerTAC simulator against non-stationary opponents (which change from one stationary strategy to another). As future work we plan to use transfer learning ideas to promote a fast learning and provide theoretical guarantees for switch detection.

REFERENCES

[1] R. I. Brafman and M. Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *The Journal of Machine Learning Research*, 3:213–231, 2003.

[2] S. P. Choi, D.-Y. Yeung, and N. L. Zhang. An environment model for nonstationary reinforcement learning. In *Advances in neural information processing systems*, pages 987–993, Denver, USA, 1999. Citeseer.

[3] P. Hernandez-Leal, E. Munoz de Cote, and L. E. Sucar. A framework for learning and planning against switching strategies in repeated games. *Connection Science*, 26(2):103–122, Mar. 2014.

[4] P. Hernandez-Leal, E. Munoz de Cote, and L. E. Sucar. Using a priori information for fast learning against non-stationary opponents. In *Advances in Artificial Intelligence – IBERAMIA 2014*, pages 536–547, Santiago de Chile, Nov. 2014.

[5] P. Hernandez-Leal, M. E. Taylor, E. Munoz de Cote, and L. E. Sucar. Bidding in Non-Stationary Energy Markets. In *Autonomous Agents and Multiagent Systems, 2015.*, Istanbul, Turkey, May 2015.

[6] W. Ketter, J. Collins, and P. Reddy. Power TAC: A competitive economic simulation of the smart grid. *Energy Economics*, 39:262–270, Sept. 2013.