On the Usefulness of Opponent Modeling: the Kuhn Poker case study

(Short Paper)

Alessandro Lazaric
Politecnico di Milano
Dept. of Elect. and Inf.
Piazza Leonardo da Vinci, 32
Milan, Italy
lazaric@elet.polimi.it

Mario Quaresimale
Politecnico di Milano
Dept. of Elect. and Inf.
Piazza Leonardo da Vinci, 32
Milan, Italy
mario.quaresimale@mail.polimi.it

Marcello Restelli
Politecnico di Milano
Dept. of Elect. and Inf.
Piazza Leonardo da Vinci, 32
Milan, Italy
restelli@elet.polimi.it

ABSTRACT

The application of reinforcement learning algorithms to Partially Observable Stochastic Games (POSG) is challenging since each agent does not have access to the whole state information and, in case of concurrent learners, the environment has non-stationary dynamics. These problems could be partially overcome if the policies followed by the other agents were known, and, for this reason, many approaches try to estimate them through the so-called opponent modeling techniques. Although many researches have been devoted to the study of the accuracy of the estimation of opponents' policies, still little attention has been deserved to understand in which situations these model estimations can be actually useful to improve the agent's performance.

This paper presents a preliminary study about the impact of using opponent modeling techniques to learn the solution of a POSG. Our main purpose is to provide a measure of the gain in performance that can be obtained by exploiting information about the policy of other agents, and how this gain is affected by the accuracy of the estimated models. Our analysis focus on a small two-agent POSG: the Kuhn Poker, a simplified version of classical poker. Three cases will be considered according to the agent knowledge about the opponent's policy: no knowledge, perfect knowledge, and imperfect knowledge. The aim is to identify which is the maximum error that can affect the model estimate without leading to a performance lower than that reachable without using opponent-modeling information. Finally, we will show how the results of this analysis can be used to improve the performance of a reinforcement-learning algorithm coped with a simple opponent modeling technique.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Learning

General Terms

Algorithms

Cite as: On the Usefulness of Opponent Modeling: the Kuhn Poker case study (Short Paper), Alessandro Lazaric, Mario Quaresimale and Marcello Restelli, *Proc. of 7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2008)*, Padgham, Parkes, Müller and Parsons (eds.), May, 12-16., 2008, Estoril, Portugal, pp. 1345-1348.

Keywords

Multi-agent Learning, Opponent Modeling, Reinforcement Learning

1. INTRODUCTION

In this paper, we propose a preliminary study about the effectiveness of using Opponent Modeling (OM) techniques to improve the performances of reinforcement-learning algorithms in Partially Observable Stochastic Games (POSG).

To this day, the main studies in this field concern the development of OM algorithms [6] devoted to improve the accuracy of opponent behavior estimation. OM approaches can be classified according to the amount of prior knowledge required for their application. Statistical classifiers, artificial neural networks, deterministic finite automata, and decision trees are examples of general-purpose methods, while expert systems, feature-based methods, and plan recognition belong to the set of domain-specific techniques. Regardless of which technique is considered, we want to point out that, when the approximation error of the estimated model is too large, using this information could prove detrimental for the learning process.

In order to avoid this eventuality, McCracken and Bowling studied OM from a different point of view, aiming to ensure efficacy from its usage. Their research is founded on admitting success of OM, but also focuses on the fact that such success depends on situations: exploiting a wrong or ineffective opponent model may drastically reduce performances. They introduced the Safe Policy Selection algorithm to profitably exploit OM [5], defining as safe a policy that leads to a total reward not lower than the expected value of the optimal policy; in this way, when OM yields to decrease such safety value, it is not used. McCracken and Bowling apply the cited above algorithm to Rock-Paper-Scissors, a zero sum matrix-game, where information about the state space is complete, while we aim to study OM effectiveness in a POSG context.

Going in the same direction of such evaluation, we provide a preliminary analysis on the performances achievable by exploiting OM techniques, in order to numerically quantify them both in worst and best cases. In particular, we show how the knowledge of the policies followed by other agents can be effectively used by the player to improve her performance. On the other hand, when the opponent's policy is not exactly known, but the player can exploit only an

estimated model based on the previously observed actions, the advantage can be significantly reduced, or it can even turn into a loss of performance. On the basis of this analysis, we experimentally show that it is possible to improve the performance of an RL agent by avoiding to exploit OM information when the accuracy of the estimated model is too low.

Recently, many research works have focused on Texas Hold'em Poker [1] [2], considered as the ideal testbed for studying POSG. Nevertheless, as Texas Hold'em is too complex for a preliminary analysis, we focus our attention on studying OM techniques in a simplified version of Poker Game: the Kuhn Poker [4]. Although this problem is quite trivial, it still has the key features of the primal game, and for this reason it was already studied, with other purposes, in past works [6] [3].

The rest of the paper is structured as follows: next section briefly describes Kuhn Poker's rules and its formalization as a POSG. In Section 3 we expose our OM analysis, which is so structured: at first, we study the case where no information about the opponent's policy is considered, then we analyze the improvement that can be obtained when the policy followed by the opponent is known, and finally we show how the use of an approximate model of the opponent's policy may have negative effects. In Section 4 we experimentally compare the performance of three RL agents: without OM, with OM, and using OM only when the model estimation is accurate enough. In the last section we draw conclusions and describe future directions.

2. KUHN POKER

Kuhn Poker is a simplified two-person poker, its rules are as follows:

- Two player, each of whom has two dollars
- 3 card deck: King (K), Queen(Q), Jack (J)
- At start, both players ante one dollar and receive a private card; the third card remains hidden to each of them.
- After anting, players can choose between two actions: BET and PASS.

After both players anting, the non-button chooses whether to BET or to PASS; then the button replies with her chosen action. A hand terminates when both players choose BET or the second action of betting sequence is PASS. The most long betting sequence is when the non-button chooses PASS and the button replies with BET: only in this case, non-button must act again, then the hand is terminated. A player wins a hand when her opponent folds, or when she has the highest card in the showdown. The game goes to showdown when both players bet or pass. If only one player bets and the other replies with a PASS, showdown does not occur. Given this betting sequence, the highest pot is 4 dollars, so the best gain an agent can obtain is 2 dollars; this occurs when both players bet. If showdown is reached by a two pass sequence, the pot is 2 dollars and the gain for the winning agent is 1 dollar.

Although the Kuhn Poker is a Partially Observable Stochastic Game (POSG), in this paper we limit our analysis to the case of fixed opponents, so that the problem can be modeled as a POMDP, that is described as the tuple: $< \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \Omega, \mathcal{O} >$, where $\mathcal{S}$ is the state space describing the environment, $\mathcal{A}$ is the set of actions that can be performed in the environment, $\mathcal{T}$ is the transition function, expressing the probability to go from a starting state to a next state when a given action is executed, and $\mathcal{R}$ is the reward function, measuring the goodness of taking an action in a certain state. Ω is the set of observations that the agent can make; $\mathcal{O} : \mathcal{S} \times \mathcal{A} \times \Omega \longrightarrow [0,1]$ is the observation function, where $\mathcal{O}(s', a, o) = P(\Omega_t = o | S_t = s', A_{t-1} = a)$ is the probability of experiencing observation o, given the performed action a and being s' the ending state. The behavior of each player is specified by her policy π, which is a function that, given a state s and an action a, returns the probability to execute a in s: $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$.

3. OPPONENT MODELING ANALYSIS

In this section, we analyze the effectiveness of exploiting information about the opponent's policy in the Kuhn Poker game. Our analysis is carried out by considering that the opponent is following a stationary policy, so that the problem can be formalized as a POMDP, where the opponent's actions can be used as observations of the hidden part of the state space. In particular, we consider opponent's policies that depend only on her private card, and, fixed one policy, we compute the utility value for the best-response policy.

Without any information about the opponent's policy, each player knows only her own private card. This means that, if a player owns a Queen, the probabilities that her opponent owns a Jack or a King are both equal to 0.5. On the other hand, by knowing the opponent's policy and observing her actions, a player can exploit this information to reduce her uncertainty about the private card of the opponent. To measure the amount of information that can be obtained about the opponent's private card by knowing her policy, we use the *mutual information* quantity between the random variable A, which represents the opponent's action, and the random variable C, which represents the opponent's private card:

$$\mathcal{I}(A; C) = \mathcal{H}(C) - \mathcal{H}(C|A), \tag{1}$$

where $\mathcal{H}$ is the entropy function, that measures the uncertainty about a stochastic variable. Since the private card of the opponent is always randomly extracted, the entropy $\mathcal{H}(C)$ is constant, and attains its maximum value. On the other hand, the conditional entropy of variable C given the value of variable A is strictly dependent from the policy π_{opp} followed by the opponent. Given the assumption that the opponent's policy depends only on the value of her private card, $\mathcal{H}(C|A)$ is formally defined as:

$$\begin{aligned} \mathcal{H}(C|A) &= -\sum_{a \in A} Pr(a) \sum_{c \in C} Pr(c|a) log(Pr(c|a)) \\ &= -\sum_{a \in A} \left(\sum_{c \in C} \pi_{opp}(c, a) \cdot Pr(c) \right) \cdot \\ &\quad \cdot \sum_{c \in C} \frac{\pi_{opp}(c, a)}{Pr(a)} log \left(\frac{\pi_{opp}(c, a)}{Pr(a)} \right). \end{aligned} \tag{2}$$

Low values of the conditional entropy $\mathcal{H}(C|A)$ (and, consequently, high values of the mutual information $\mathcal{I}(A; C)$) mean that, by knowing the opponent model and observing her actions, we can significantly reduce the uncertainty

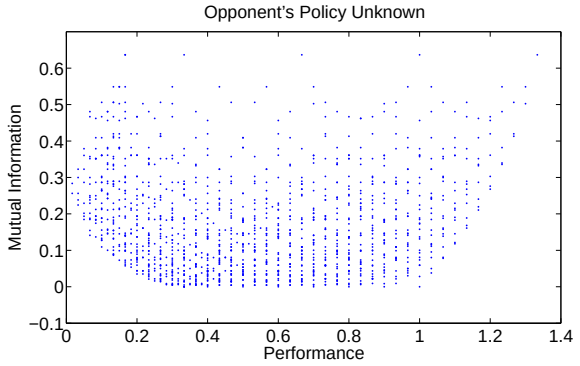

Figure 1: Mutual information and best-response performance for 1,000 fixed opponent's policies. No information about the policies is used.

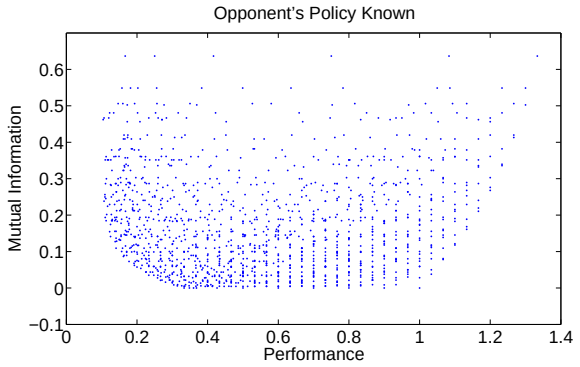

Figure 2: Mutual information and best-response performance for 1,000 fixed opponent's policies. Policies are exactly known.

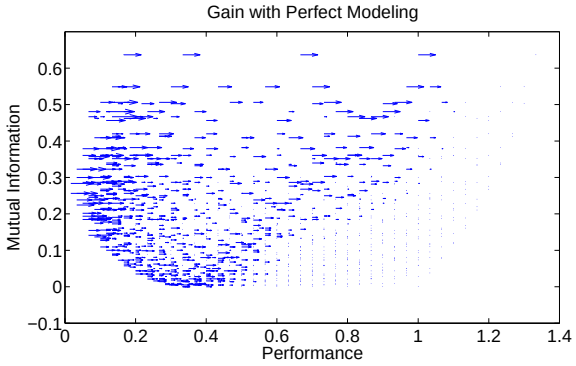

Figure 3: The arrows show the improvement due to the knowledge of the opponent's policy.

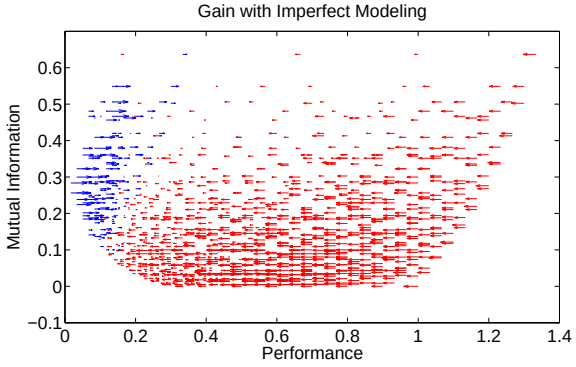

Figure 4: The arrows show the improvement (blue) or the worsening (red) when the player exploits an estimated model whose distance from the actual opponent's policy is up to 0.1.

about the opponent's card. On the other hand, when the opponent follows a policy that does not depend on the value of her own card (e.g., a random policy), the conditional entropy $\mathcal{H}(C|A)$ is equal to the entropy $H(C)$, so that the mutual information is zero; in these cases, the use of OM techniques is useless.

To study the effect of OM techniques in the Kuhn Poker, we consider several possible stationary policies for the opponent. For each one of these policies, we compute the corresponding mutual information $\mathcal{I}(A;C)$ and the performance attained by its best-response policy. In general, given an opponent policy π_{opp}, the expected performance of a policy π is the average of the expected values of the states weighted by the probability of visiting the corresponding state:

$$U(\pi|\pi_{opp}) = \sum_{s \in \mathcal{S}} Pr(s|\pi, \pi_{opp}) V(s|\pi, \pi_{opp}).$$

The best-response policy π^* against a given policy π_{opp} is the one which attains the highest utility:

$$\pi^*_{\pi_{opp}} = arg \max_{\pi \in \Pi} U(\pi|\pi_{opp}).$$

Figure 1 shows relation between the mutual information of the opponent's policy (y-axis) and the performance of its best-response policy (x-axis), in the case where no information about the opponent's policy is exploited [1]. As we can see from the graph, the opponent's policy which is less exploited has a quite high mutual information value. In particular, it is worth noting that there is no policy for the opponent that is placed near the origin of the graph. This means that, if the opponent wants to adopt a policy that can be hardly exploited, it has to follow a policy that reveals information about her private card. On the other hand, when the opponent wants to hide at most the value of her card, she can be easily exploited. This trade-off is what makes the use of OM techniques interesting.

Figure 2 shows how the situation changes when the player knows the policy followed by the opponent, so that the problem can be formalized as a POMDP and solved by using the observable histories as state information. As it can be noticed, several points have been moved to the right, since exploiting the information of the opponent's policy has allowed to identify best-response policy with higher performances. To give a better visualization of the effect of knowing the opponent's policy, in Figure 3 we have used arrows to represent the gain. As expected, when the mutual information is low, knowing the opponent's policy results in small gains.

[1]Each point corresponds to a different opponent's policy. The 1,000 policies have been generated by considering ten evenly-spaced values for the probability of betting given each value of the private card.

On the other hand, it is not always true that the knowledge of policies that convey much information can lead to high gains, especially when the policies are quite weak (look at the right side of the graph).

Unfortunately, in adversarial problems, a player does not know the policy of her opponent. For this reason, the resort to OM techniques is quite common. The problem is that the estimated model is an approximation of the policy actually followed by the opponent. Using a model affected by a large approximation error could lead to a performance that is worse than that achievable using no model at all. Figure 4 shows how much, in the worst case, the performances change when the distance between the actual opponent's policy and the estimated one is not larger than 0.1[2]. The arrows that point toward left (red arrows) corresponds to opponent's policies for which the agent may have a loss of performance when using a model with a low accuracy. As we can notice, the opponent's policies that are associated to larger losses are those that have little or not advantage when knowing the actual policy followed by the opponent.

In the next section, we show how this analysis can be used to improve the performance of an RL player.

4. LEARNING EXPERIMENTS

In this section, we show some preliminary experiments obtained by using Q-learning [7], a popular reinforcement-learning algorithm, against a fixed opponent. In particular, we consider three different versions of Q-learning:

- Q-learning without OM: the state space depends only by the player's private card;
- Q-learning with OM: the state space depends by the player's private card and by the observed opponent's action[3]
- Q-learning with reliable OM: in this version, we keep an estimate of the accuracy of the opponent's model, so that when the accuracy is below a certain threshold we use Q-learning without OM, otherwise we exploit the opponent modeling information. The choice of the threshold is made according to the analysis described in the previous section.

In Figure 5, the performances of the three learning algorithms are represented. As it can be noticed, Q-learning with OM is ineffective in the first learning steps when the information about the opponent's policy is still highly uncertain. On the other hand, Q-learning without OM is able to quickly learn a good solution, but it has not enough information to exploit the opponent at best. As we can notice, the third approach, which uses the opponent-modeling information only when it is accurate enough, is able to attain both a good learning speed and a good performance in the long-run.

5. CONCLUSIONS

[2]The distance between two policies is computed as the L2-norm of the difference vector between the two vectors that specify the policies in the two models.

[3]In this problem, this information is equal to the history of observations, thus allowing Q-learning to solve the POMDP problem.

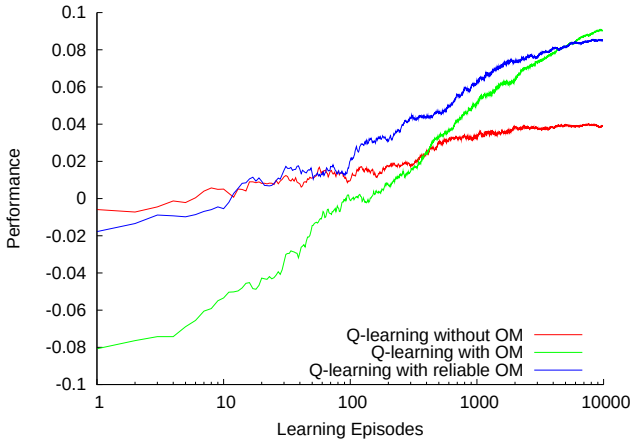

Figure 5: Comparison of three RL algorithms against a fixed opponent's policy. Results are averaged over 1,000 runs

In this paper we have presented a preliminary study on measuring the usefulness of using opponent-modeling techniques in Partially Observable Stochastic Games, by focusing on a simple poker game. The results of our analysis show that, in a context like Kuhn Poker, OM technique can be very useful, but only under the necessary condition that the model describing the opponent's behavior is accurately estimated.

This paper represents just a first step and opens several directions for future research. The following steps will be devoted to extend this analysis to cases where the opponent can adopt more complex policies, such as stationary policies that consider the actions performed by the player, non-stationary policies, and non-stationary policies based on OM information. The final goal of this work is to extend the results of these analyses to more complex problems, such as the Texas Hold'em Poker.

6. REFERENCES

[1] D. Billings, A. Davidson, J. Schaeffer, and D. Szafron. The challenge of poker. *Artificial Intelligence*, 134(1-2):201–240, 2002.

[2] D. Billings, D. Papp, J. Schaeffer, and D. Szafron. Opponent modeling in Poker. In *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-98)*, pages 493–498, Madison, WI, 1998. AAAI Press.

[3] D. Koller and A. Pfeffer. Representations and solutions for game-theoretic problems. *Artificial Intelligence*, 94(1-2):167–215, 1999.

[4] H. Kuhn. A simplified two person poker. In W. H. Kuhn and A. W. Tucker, editors, *Contributions to theory of games*, pages 97–103. Princeton University Press, 1950.

[5] P. McCracken and M. Bowling. Safe strategies for agent modelling in games. In *AAAI 2004 Symposium on Artificial Multi-Agent Learning*. AAAI Press, 2004.

[6] T. Schauenberg. Opponent Modeling and Search in poker. Master's thesis, University of Alberta, 2006.

[7] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. Bradford Book, 1998.