An Alternative Approach to the Revision of Ordinal Conditional Functions in the Context of Multi-Valued Logic

Klaus Häming and Gabriele Peters

University of Applied Sciences and Arts,
Computer Science, Visual Computing,
Emil-Figge-Str. 42, D-44221 Dortmund, Germany

Abstract. We discuss the use of Ordinal Conditional Functions (OCF) in the context of Reinforcement Learning while introducing a new revision operator for conditional information. The proposed method is compared to the state-of-the-art method in a small Reinforcement Learning application with added futile information, where generalization proves to be advantageous.

1 Introduction

An autonomous learning system tries to figure out which actions are beneficial and which have to be avoided. Starting with three system requirements we developed the work described in this paper. These requirements are the following.

First, an autonomous learning system should be able to learn from experience. It should have some kind of memory that, e.g., enables it to decide not to fall over a cliff again in case it proved harmful the last time. A widely adopted approach to incorporate such a property is given by Reinforcement Learning (RL) [10].

Second, the system should generate a representation of its belief that allows further reasoning. In this area Belief Revision (BR) techniques can be found. We will examine the usefulness of Ordinal Conditional Functions (OCF) [4,7] in this work.

Third, and most important, we want both mentioned approaches to benefit from each other. This kind of mixture of low-level learning-by-doing and high level deduction abilities is called a two-level learning approach. Psychological findings [6] indicate that such two-level learning principles can explain some of the human learning abilities. While humans are able to learn both, in a top-down and bottom-up way[9], we will focus on the bottom-up part only.

A combination of RL and BR has been proposed recently [5], influenced by [8] and [11]. In this work, we will shed light onto a rather small but important detail of this approach.

2 Reinforcement Learning and OCF

In Reinforcement Learning, we have a set of states $\mathfrak{S}$, a set of actions $\mathfrak{A}$, a transition function $\delta : \mathfrak{S} \times \mathfrak{A} \to \mathfrak{S}$, and a reward function $r : \mathfrak{S} \times \mathfrak{A} \to \mathbb{R}$. Belief

K. Diamantaras, W. Duch, L.S. Iliadis (Eds.): ICANN 2010, Part II, LNCS 6353, pp. 200–203, 2010.

about good and poor actions is learned by applying a learning scheme, in this case we use Q-learning. The experience is captured in the Q(uality)-function, that assigns an expected reward to each state-action-pair. One can interpret the Q-function in such a way, that an action A is *believed* to be best, if it has the highest $Q(S, A)$ value for a given state S. This is where we establish a connection to the high-level belief using BR.

BR is a theory of maintaining a belief base in such a way, that the current belief is reflected in a consistent manner (cf. [2] and [1]). We model our belief base κ as an OCF. This is a ranking function that maintains a list of all models, which are propositional information in the form of conjunctions. The models the system believes in are set to rank 0, while all ranks greater than 0 reflect an increasing disbelief. We denote the rank an OCF κ assigns to a model M as $\kappa(M)$. By convention, contradictions shall have the rank ∞.

However, during the exploration the information gathered and the information needed is in the form of conditionals, not conjunctions. To check, whether an OCF believes a conditional, it is sufficient to compute the belief ranks $r_1 = \kappa(SA)$ and $r_2 = \kappa(S\overline{A})$. If $r_1 < r_2$, the conditional is believed.

More difficult then querying the belief base, is its update, called *revision*. The revision operator is "$*$".

Conditionals in BR are usually denoted as $(A|S)$, where S is the antecedent and A the consequent. The meaning of $(A|S)$ is not exactly the same as $S \Rightarrow A$ [4]. The latter means that S implicates A irrespective of the values of other variables. In contrast, $(A|S)$ expresses that A is believed if κ is conditioned with S and S alone. In contrast, a revision $(\kappa * (ST))$ may not result in A being believed.

The revision described in [5], conforms to $(\kappa * (A|S))$. The new revision we introduce is $(\kappa * (S \Rightarrow A))$. It needs a new operator $\kappa[A]$, which shall return the highest disbelief among all models of A. $(\kappa * (S \Rightarrow A))$ is defined as follows:

If $\kappa(SA) < \kappa(S\overline{A})$, do nothing. If $\kappa(SA) \geq \kappa(S\overline{A})$, then the OCF κ' derived from κ by rearranging the models using

$$\begin{aligned} &\forall M \in \mathfrak{M} : \kappa'(M) := (\kappa * (S \Rightarrow A))(M) \\ &= \begin{cases} \kappa(M) - \kappa(S \Rightarrow A), \text{ if } M \text{ is a model of } S \Rightarrow A \\ \kappa[SA] + 1 - \kappa(S \Rightarrow A) + \kappa(M) - \kappa(S\overline{A}), \text{ if } M \text{ is a model of } S\overline{A} \end{cases} \end{aligned} \tag{1}$$

will result in $\kappa'(SA) < \kappa'(S\overline{A})$. Consequently, κ' expresses the belief in $S \Rightarrow A$.

Concerning the insurance of actual belief, this method works just as good as $(\kappa * (A|S))$, but introduces greater changes. The justification for these changes is its behavior toward sequences of belief changes, especially in the context of multi-valued logic, where $(\kappa * (A|S))$ fails to produce consistent results when considering negation and generalization.

3 Application

We examine the effect of the proposed algorithm in a cliff-walk gridworld [10] (Figure 2). For this application, three cases are examined, which are plain

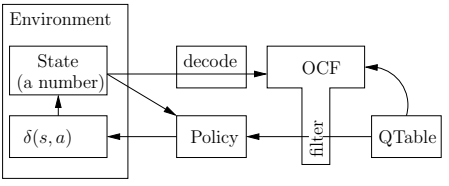

Fig. 1. OCF-augmented RL system. The OCF acts as a filter that limits the choices of the policy.

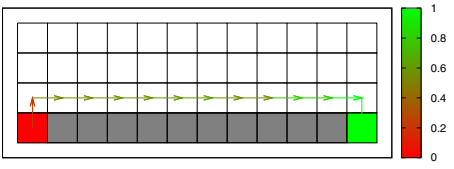

Fig. 2. Cliff-walk gridworld. The goal of a moving agent is to reach the green square, starting from the red one. Entering the dark squares (representing a cliff) results in a high negative reward. Superimposed is the learned path after 100 episodes. The path color indicates the expected reward by displaying the value of $\min(1, \frac{\text{expected reward}}{\text{goal reward}})$ using the displayed color key.

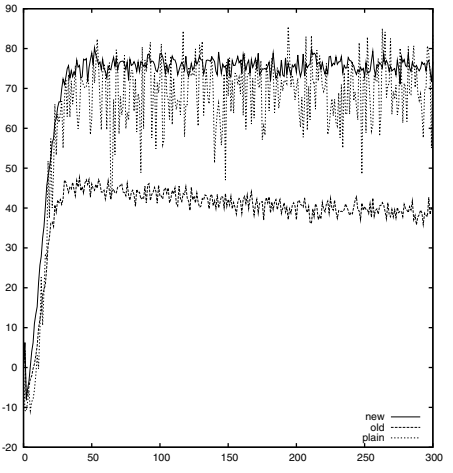

Fig. 3. Results. The diagrams show the rewards over a series of 300 episodes. *plain* shows the result of a plain Q-learner, *old* shows the result of revisions with $(\kappa * (A|S))$, and *new* shows the result of revisions with $(\kappa * (S \Rightarrow A))$. The values are averages of 1000 runs.

Q-learning, OCF-augmented Q-learning with application of $(\kappa * (A|S))$, and OCF-augmented Q-learning with application of $(\kappa * (S \Rightarrow A))$. An OCF-augmented Q-learner is a Q-learner that has conditionals extracted from its Q-Table. These conditionals revise the learner's OCF and this OCF acts as a filter for the choice of actions afterwards. Figure 1 shows this architecture.

We add futile information to model the case where the agent perceives properties of its environment that are not helpful with regard to its goal. The OCF-augmented Q-learners are expected to be able to generalize and therefore identify the futile information. The generalization is performed in the same way as in [5] by counting the pattern frequency. The general idea is to keep track of how often sub-patterns of antecedents are used in the context of particular consequents. If they occur frequently enough, we revise the OCF with the sub-pattern instead of the complete state description. The state description is also adopted from [5], where a qualitative description is used which consists of the relative position of the agent towards the goal (north, south, east, west) and a distance (near, middle, far) amended with information on adjacent obstacles. Reaching the goal triggers a reward of 100, getting closer towards it is rewarded with 0.5. Stepping into the chasm is punished by -10, every other step gets a -1. After 100 steps the episode is forced to end. The results are depicted in Figure 3. It is evident that a revision with $(\kappa * (S \Rightarrow A))$clearly outperforms a revision with $(\kappa * (A|S))$.

The latter is worse than a plain Q-learner and even seems to deteriorate over time. An explanation for this may lie in the fact that the OCF gets contaminated by harmful conditionals. However, this has not been examined in this work.

4 Conclusion

There are some questions left open. Clearly, the use of an OCF speeds up the learning process (measured in the number of episodes). However, the role of futile information has to be examined in more detail. The performance of the proposed method surpasses the plain Q-learner's. Since off-policy learning usually shows a worse performance than on-policy learning, OCF-augmentation could be a way to ease this weakness. Finally, it may be interesting to examine the use of an OCF directly as a Q-function to create a Relational Reinforcement Learning system[3].

Acknowledgments. This research was funded by the German Research Association (DFG) under Grant PE 887/3-3.

References

1. Alchourron, C.E., Gardenfors, P., Makinson, D.: On the logic of theory change: Partial meet contraction and revision functions. J. Symbolic Logic 50(2), 510–530 (1985)
2. Darwiche, A., Pearl, J.: On the logic of iterated belief revision. Artificial intelligence 89, 1–29 (1996)
3. Dzeroski, S., De Raedt, L., Driessens, K.: Relational reinforcement learning. Machine Learning 43, 7–52 (2001)
4. Kern-Isberner, G.: Conditionals in nonmonotonic reasoning and belief revision: considering conditionals as agents. Springer, New York (2001)
5. Leopold, T., Kern Isberner, G., Peters, G.: Combining reinforcement learning and belief revision: A learning system for active vision. In: BMVC 2008, pp. 473–482 (2008)
6. Reber, A.S.: Implicit learning and tacit knowledge. Journal of Experimental Psycology: General 3(118), 219–235 (1989)
7. Spohn, W.: Ordinal conditional functions: A dynamic theory of epistemic states. In: Causation in Decision, Belief Change and Statistics, pp. 105–134 (August 1988)
8. Sun, R., Merrill, E., Peterson, T.: From implicit skills to explicit knowledge: A bottom-up model of skill learning. Cognitive Science 25, 203–244 (2001)
9. Sun, R., Zhang, X., Slusarz, P., Mathews, R.: The interaction of implicit learning, explicit hypothesis testing, and implicit-to-explicit knowledge extraction. Neural Networks 1(20), 34–47 (2006)
10. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (1998)
11. Ye, C., Yung, N.H.C., Wang, D.: A fuzzy controller with supervised learning assisted reinforcement learning algorithm for obstacle avoidance. IEEE Transactions on Systems, Man, and Cybernetics, Part B 33(1), 17–27 (2003)