Contribution to the study and the design of reinforcement functions

by Juan Miguel Santos
Directed by Dr. Hugo Scolnik and Norbert Giambiasi
April, 12nd, 1999, Universidad de Buenos Aires

The underlying concept in Reinforcement Learning is as simple as it is attractive: to learn by trial and error from the interaction with the environment. This approach allows us to deal with problems where a learning technique searches to improve the performance of the agent (the learner) over time. Reinforcement Learning groups a set of such techniques, and it uses a performance measure based on two types of signals given by a Critic or Reinforcement Function: penalty and reward.

The use of these techniques is closely related to the conditions imposed on each type of problem. In the easiest case, a problem is provided with a world model (Reinforcement and Transition Functions) and can be modeled by means of a Markovian Decision Process. In more difficult cases, the Transition Function is unknown and, today, two widely known algorithms based on the Temporal Differences Method: Adaptive Heuristic Critic [Sutton, 1988] and Q-Learning [Watkins, 1989] are able to deal with the situation of an unknown model.

If the state-action space of the system is so large that it forbids an explicit representation of the agent's internal structure (i.e. tables, tuples, etc.), it is mandatory to consider strategies that generalize. Today, the state of the art positions Artificial Neural Networks as an obvious tool to implement generalization. Moreover, large spaces complicate the exploratory process. In particular, the exploration-exploitation dilemma becomes especially delicate. Therefore, it is necessary to develop exploration strategies closely related to those of memorization and generalization.

Over the last ten years, the evolution of the Reinforcement Learning techniques has impacted on a closely related field: Robot Learning. In Robot Learning, a robot must improve its performance over time; and the RL paradigm suits this definition well in that the performance measure can be used to achieve the goal: In fact, the Discount Reinforcement Infinite Sum is used due to its useful convergence property.

However whatever criterion is used, in all cases it depends of the Reinforcement Function definition. Unfortunately, it is not simple to define such a function, especially in many applications with physical systems (i.e., robots) whose states are represented by means of the sensor activity (numerical values) where humans are more at ease with symbolical values. Thus, the definition of the Reinforcement Function constitutes a critical point for the implementation of RL.

The objective of this thesis is to study the design of the Reinforcement Function and introduce an approach to undertaking this task. Additionally, we study the feasibility and consequences of adapting the Reinforcement Function during the learning process so as to improve the performance of the system.

As a first step, we propose a general expression for the RF. It is expressed in terms of constraints, which depend of a set of values: the Reinforcement Function Parameters. We also suggest a Reinforcement Function Design Process, with two main stages. The first one translates a natural language description into an instance of the Reinforcement Function General Expression. The second one tunes parameters of constraints for obtaining the optimal definition of the function (relative to exploration).

Based on a particular (but generic) case of the Reinforcement Function with two constraints (one associated with positive reinforcement and the other with negative ones), we propose an analytic method and an algorithm (Update Parameters Algorithm, or UPA) to obtain the value of the Reinforcement Function Parameters.

As a second step, we studied the possibility and utility of changing the definition of the Reinforcement Function (parameter tuning) in a dynamic way during the Learning Phase, and we show the consequences on the exploration- exploitation dilemma.

A rule for tuning the RF parameters during UPA execution involves changing the function to obtain an ideal ratio of positive and negative reinforcements defined in advance. Before learning, UPA guarantees a balanced proportion of reinforcements during exploration. During learning, UPA modulates the exploration-exploitation relation. In both cases, the results are improvements of the learning performance (quality and time).

The idea of using the variations of the performance as an evaluation criterion has been partially explored by some authors. For example, [Schmidhuber et al., 1997] use as a criterion acceptance (or rejection) of previously imposed changes on a Policy, the evolution of the Reinforcements Sum received from the last change. [Mataric, 1994], and [Millán, 1996] use progress estimators to measure the evolution of the system in achieving sub-objectives (for simple behaviors with an associated metric) expressed in the RF. On the other hand, [Ackley and Littman, 1991] introduce an Evolutionary Reinforcement Learning scheme where RL allows individuals to improve their performance along their life. Thus, the performance of the complete system is indirectly used for accepting, or not, changes of the RF definition of each individual. However, it is our opinion that these attempts failed to achieve a rigorous development in regards to the exploration-exploitation dilemma, and its related matter: learning speed, convergence, etc.

The neural implementation of the RL used in the experimental parts of this thesis, is based on the clustering properties of the Radial Basis Function Artificial Neural Networks [Moody and Darken, 1989]. The need to achieve an appropriate discretization of the situation-action space through clustering has lead us to add to the update rule of the neural network a growing strategy.

In this dissertation we will review the Reinforcement Learning and the associated open issues. After that, we will describe our contribution to the study and development of the Reinforcement Function. Then, we will illustrate, with several experiments involving robots (mobile and manipulator), the efficiency of our proposed method. Results will be analyzed and discussed. Finally, we will emphasize the main conclusions of this work and present some future directions of research.

References

1. David **Ackley** and Michael **Littman**, **1991**. Interactions Between Learning and Evolution, in Artificial Life II, SFI Studies in the Sciences of Complexity, vol. X, eds. By C.G. Langton, C. Taylor, J. D. Farmer & S. Rasmussen, Addison Wesley Publishers, 487-509
2. Maja J. **Mataric**, **1994**. Reward Functions for Accelerated Learning, in *Machine Learning: Proceedings of the Eleventh International Conference*, William W. Cohen and Haym Hirsh, eds. Morgan Kaufmann Publishers, 181-189.
3. J. del R. **Millán**, **1996**. Rapid, Safe and Incremental Learning of Navigation Strategies, *Special Issue on Learning Autonomous Robots, M. Dorigo Guest Editor, IEEE Trans. on Systems, Man and Cybernetics - part B*, Vol. 26, No. 3, 408-420.
4. John **Moody** and Christian J. **Darken**, **1989**. Fast Learning in Networks of Locally-Tuned Processing Units, *Neural Computation,* **1**, 281-294.
5. Jürgen **Schmidhuber**, Jieyu **Zhao**, Nicol N. **Schraudolph**, **1997**. Reinforcement Learning with Self-Modifying Policies, in Learning to Learn, Ed. by S.Thrun and L. Pratt, Kluwer Publishers, 293-309.
6. Richard S. **Sutton**, **1988**. Learning to Predict by the Methods of Temporal Differences, *Machine Learning,* **3**, 9-44
7. Christopher John Cornish Hellaby **Watkins**, **1989**. Learning from delayed rewards, Ph.D. Thesis, King's College, University Library Cambridge.