Modified Reinforcement Learning Infrastructure

Jyrki Suomala[1] and Ville Suomala[2]
[1]NeuroLab, Laurea University of Applied Sciences, Vanha maantie 9, FI-02650, Espoo, FINLAND
[2]University of Oulu, P.O Box 3000, FI-90014, University of Oulu, FINLAND
Jyrki.suomala@laurea.fi, ville.suomala@oulu.fi

Abstract—The reinforcement learning (RL) model has been very successful in behavioral sciences, artificial intelligence and neuroscience. Despite its fruitfulness in many simple situations, the RL model does not always cope well with real life situations involving a large space of possible world states or a large set of possible actions. We propose a modified version of the RL learning model. The benefit of this model is that the temporal difference prediction error can be used directly to update not only the value of the latest action of the learning agent, but the values of many possible future actions. An example application of this modified reinforcement learning infrastructure (MRLI) is presented for a customer behaviour in a complex shopping environment.
Index Terms – MRLI, Graph Theory, Learning, Behavioral model, Decision-making.

1. Introduction

Human behavior like any other complex dynamic system needs some kind of rules in order to organize constantly flowing information. In earlier research, the goal has been to understand neurobiological background of human behavior by analyzing as simple units of behavior as possible. More recently, many models aim to analyze human behavior from its goal point of view, when the process is very complex [1], [2]. Reinforcement learning (RL) model has been very successful in behavioral sciences, artificial intelligence and neuroscience. This model can describe a learning mechanism based on prediction error signals, which measure the discrepancies between actual and expected outcomes [3], [4]. The RL theories in neuroscience assume that an agent learns state-action values by a trial and error procedure, and these values are then used as decision variables to guide choice [5].

Ideas from RL have been applied to explain a wide range of behavioral phenomena both in behavioral and neurophysiological level [6]. In particular, correlates of prediction errors in the striatum have been found in multiple neurophysiological studies [7]–[9].

Despite the fruitfulness of the RL model in many areas it yet has many limitations, the scaling problem being one of the biggest challenges [3], [6], [10]. The basic RL model does not cope well with domains involving a large space of possible world states or a large set of possible actions. Therefore, most of the RL models have been applied to highly simplified learning situations.

The hierarchical reinforcement learning (HRL) model is one of the newest attempts to describe human behavior in more complex situations [6]. Whereas the standard RL model allows an agent to select among primitive actions, the HRL model lets an agent also to select subroutines, each associated with its own behavioral policy and its own designated subgoals [6]. In the HRL model, actions cohere into subtask sequences, which fit together to achieve overall goals. Despite the fact that the HRL model can partially solve the scaling problem, which has been the basic problem of the standard RL models, it has still many restrictions. In particular, the HRL cannot be directly applied in situations, where there are strong correlations between different actions that are far away in the space of locations. The current paper presents a modified reinforcement learning infrastructure (MRLI), which aims to solve one of the problems of the HRL; in the current model, the temporal difference (TD) prediction error can be used directly to update not only the value of the latest action, but the values of many possible future actions. In this way, more realistic behavior and learning process of a human agent could be described. For example, a consumer who is intended to buy a new tablet computer, could make her decision after seeing only one ad on the internet. However, some other consumer could only decide after tens of advertisements, visits to online stores and so on. The MRLI model can be used to describe both kind of behavior.

2. MRLI model

Graph theory provides a realistic framework for investigating complex networks such as human behavior, cortical networks and traffic [11]. It is also a suitable tool in analyzing complex human behavior (shopping, decision-making), in which 'touchpoints' are nodes and the paths between the nodes are edges. It is convenient to describe feedback and feedforward pathways between states using graph theoretical models. In standard RL models the driving force of an agent behavior depends on value function that equals the average sum of all future rewards received up until the end of the learning process [12]. However, it would be practically difficult to make good estimates based on the sum of all future reward at specific state. This is because the agent needs to wait until all rewards are received at the end of the learning process [12]. In the HRL [3] as well as in the current model, it is possible to count and update the values of the actions at every time step and at every state of the learning process. On empirical level, the total value of the actions in the system could be measured.

In the following, we propose a theoretical model of reinforcement learning. Our model is an extension of the reinforcement learning presented in Appendix A of the reference [3]. For simplicity, we present here a non-hierarchical version of the model, but a HRL version of the model can be obtained by modifying the setting in a straightforward manner as in [3].

The main difference between the present setting and some other RL models is, that in the current model the temporal

difference prediction error can be used directly to update not only the value of the latest action, but also the values of actions available in other locations of the state space.

Thus an agent does not learn the values of actions individually, rather an agent must discover how the information at the current moment connects up with systems of actions available in the whole system. The predictive value of a single action is assessed against an entire system of information-outcome relationships [13], [14].

The model consists of two components. In mathematical terms, the first component is a directed multigraph $G = (S,A)$, where S is the vertex set of the multigraph and denotes all possible `locations' of the agent and A is a set of directed edges between the elements of S. Each edge $a \in A$ has its source node $s_0(a) \in S$ as well as target $s_1(a) \in S$. The purpose of A is to denote all possible actions of an agent acting on the system. Given $s \in S$, the set

$$A(s) := \{a \in A | s_0(a) = s\}$$

is the set of all possible actions available at location s. If an action $a \in A(s)$ is selected, the agent moves from the location $s_0(a)$ to a new location $s_1(a)$. Note that loops are allowed so that it is possible that $s_0(a)=s_1(a)$. This means that choosing the action a does not change the location of the agent.

We associate a parameter $r(a)$ to each element $a \in A$. This is used to denote the reward gained after completing the action a. For simplicity, we assume that $r(a)$ is deterministic, but in general the reward function could also change in time.

The other component consists of a collection of action values $\{V(a)\}_{a\in A}$ and action weights $\{W(a)\}_{a\in A}$ maintained by the agent. The values of $V(a)$ and $W(a)$ change in time according to the learning strategy of the agent. The agent performs a random walk in S by following the edges of the multigraph, i.e. choosing at location s one of the possible actions $a \in A(s)$. If the agent is at location s, given the values $V(a')$ and weights $W(a')$ for $a' \in A(s)$, the agent chooses an action $a \in A(s)$ with probability

$$\mathrm{P}(a) = \frac{\exp(\frac{W(s,a)}{\tau})}{\sum_{a' \in A(s)} \exp(\frac{W(s,a')}{\tau})} \qquad (1)$$

where $0 < \tau < +\infty$ is a temperature parameter (see the Remark below).

The key point in the model is how the values of $V(a)$ and $W(a)$ are updated as the agent explores the system. Suppose at time t, the agent is at the location s_t and the values and weights are $V_t(a')$, $W_t(a')$, $a' \in A$. If the agent selects the action $a \in S(a)$, this yields the TD-prediction error

$$\delta = r(a) - V_t(a).$$

After this, action values and weights are updated for all $a' \in A$ using the equations

$$\begin{aligned} V_{t+1} &= V_t(a') + \delta\alpha_V(a,a') \qquad (2) \\ W_{t+1} &= W_t(a') + \delta\alpha_W(a,a'), \end{aligned}$$

where $\alpha_V(a,a')$, $\alpha_W(a,a')$ are (deterministic) learning rate parameters reflecting possible correlations between the actions a,a' (see the Remark below).

Now that all parts of the model have been defined, the learning process can be described as follows: In the beginning the initial location s_0, and the numbers $V_0(a')$ and $W_0(a')$ are fixed either deterministically or randomly and the action $a \in A(s_0)$ resulting to a location $s_1=s_1(a)$ is then stochastically determined using the equation (1). Using the TD-prediction error, the values of $W(a')$ and $V(a')$ are then updated to $W_1(a')$ and $V_1(a')$ as in equation (2). The next action $a \in A(s_1)$ leading to a new location s_2 is then selected using (1) with the updated values of the V:s and W:s and so on. In practical situations, a number of goal locations (often just one) are defined and the process is terminated once any of these locations is reached.

Remarks concerning the model. In some related models, S is called the state space. We call the elements of S locations since the state of the system at time t refers not only to the value of s_t, but also to the action values and weights at time t.

If the value of the parameter τ in (1) is large, the agent is more likely to explore actions a with small action value, whereas if τ is close to zero, the agent chooses actions with largest action values with high probability.

Large value of $\alpha(a,a')$ in (2) means that the TD-prediction error gained by performing the action a strongly correlates with the probability of choosing the action a' in the future. Whereas if $\alpha_V(a,a')$, $\alpha_W(a,a')$ are zero (resp. small), then the result of action a has no (significant) effect on the probability of choosing a' later on.

3. An Application of MRLI to consumer journey

In a real market place, a lot of different stimuli are seen; other people, advertisements, products and the whole marketing environment. The growing technological development has made the marketing environment much more complicated. While consumers are exposed to an expanding fragmented array of marketing touchpoints across the media, the sales channels selling process has changed from sequential steps to a process that is largely nonsequental [15]. For example, when a consumer views a TV spot for a new version of a tablet computer, she can use her current mobile device to search for more information. Then pops up a paid search link for a new device and she has access to various reviews. When the consumer reads a review, she notices a display ad from a local seller, but decides not to click it. One of the reviews contains a link to online videos people have made about their new tablets. When viewing some of the video clips she finds also other ads from different brands. During her commute to work she realizes ads on a billboard she has not seen before and then receives a direct-mail piece from a company offering a time-limited deal. She visits a local dealerships websites and finally decides to visit a Flagship store to buy the devise [15].

The example above describes how a consumer meets many marketing touchpoints before she visits a concrete shopping center or decides to buy online. In order to account the potential influence of all the stimuli experienced during the consumer decision process, the concept of experienced utility has been used [16]. This concept refers for consumer's pleasure and displeasure at each moment of experience [16], [17]. It is possible to model such situations by the MRLI model by measuring consumer's experienced

utility – behaviorally and neurophysiologically - at each moment during the whole shopping process. Here the locations s ϵ S correspond to the various marketing touchpoints. The samples in this process could have only two steps or tens or even hundreds of steps.

To set up a real test situation, the TD-prediction errors are measured to collect data that can be then used to estimate both the rewards r(a) gained from different actions a as well as the correlation parameters $\alpha_V(a,a')$, $\alpha_W(a,a')$ for different action pairs (a,a'). After this, the MRLI-model can be used to predict the typical behavior of an agent acting on the system.

4. Conclusions

Motivated by real life decision making situations, such as consumer behavior in a complex shopping environment, we have proposed a modified reinforcement learning model. The main advantage in the model, is the possibility to describe correlations between various actions available for the agent in different locations. The model therefore provides the possibility that it can account for a broader range of data than the previous RL and HRL –models. This makes the model more realistic for concrete measurements and applications.

The MRLI model is presented here on a theoretical level. The next step will involve testing the model on empirical level with human subjects and also by using computer simulations. This way, it is possible to decide, whether our model has more explanatory power than the other RL and HRL –models. According to the HRL model, brain computes two prediction error signals during the learning process. The global prediction error signal guides policy updating, whereas pseudo prediction error signals guides subroutine learning [5].

While it has been shown that the striatum nuclei produce prediction errors at subroutine level [5], it is still unclear how the prediction error signals at different hierarchical learning levels interacts in the human brain in complex situations. By organizing experiments using the MRLI it is possible to test whether there are some other areas in the brain, which combine prediction error signals from all hierarchical levels or whether the striatum have some topographical organizations that could support a coarse hierarchy [5]. The MRLI model presented here is not, as yet, so deeply developed in terms of its behavioral and neurophysiological underpinning. However, it provides a promising theoretical foundation for investigating these issues both behaviorally and neurophysiologically in the complex human choice environments. Of particular interest, are experiments in which separate prediction error signals for goals and subgoals in a shopping process will be investigated by using the MRLI approach.

Acknowledgment

The research of J. Suomala was supported by the NeuroService-project funded by the Technology Agency of Finland.

References

[1] P. W. Glimcher, *Decisions, Uncertainty, and the Brain: The Science of Neuroeconomics*, 1 edition. Cambridge, Mass: The MIT Press, 2003.

[2] D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY, USA: Henry Holt and Co., Inc., 1982.

[3] M. M. Botvinick, Y. Niv, and A. C. Barto, "Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective," *Cognition*, vol. 113, no. 3, pp. 262–280, Dec. 2009.

[4] Y. Niv and G. Schoenbaum, "Dialogues on prediction errors," *Trends Cogn. Sci. (Regul. Ed.)*, vol. 12, no. 7, pp. 265–272, Jul. 2008.

[5] N. D. Daw, "Chapter 16 - Advanced Reinforcement Learning," in *Neuroeconomics (Second Edition)*, P. W. Glimcher and E. Fehr, Eds. San Diego: Academic Press, 2014, pp. 299–320.

[6] M. M. Botvinick, "Hierarchical reinforcement learning and decision making," *Current Opinion in Neurobiology*, vol. 22, no. 6, pp. 956–962, Dec. 2012.

[7] J. Li, S. M. McClure, B. King-Casas, and P. Read Montague, "Policy Adjustment in a Dynamic Economic Game," *PLoS ONE*, vol. 1, no. 1, p. e103, Dec. 2006.

[8] W. Schultz, P. Dayan, and P. R. Montague, "A Neural Substrate of Prediction and Reward," *Science*, vol. 275, no. 5306, pp. 1593–1599, Mar. 1997.

[9] T. Schönberg, N. D. Daw, D. Joel, and J. P. O'Doherty, "Reinforcement Learning Signals in the Human Striatum Distinguish Learners from Nonlearners during Reward-Based Decision Making," *J. Neurosci.*, vol. 27, no. 47, pp. 12860–12867, Nov. 2007.

[10] J. J. F. Ribas-Fernandes, A. Solway, C. Diuk, J. T. McGuire, A. G. Barto, Y. Niv, and M. M. Botvinick, "A Neural Signature of Hierarchical Reinforcement Learning," *Neuron*, vol. 71, no. 2, pp. 370–379, Jul. 2011.

[11] N. T. Markov, M. Ercsey-Ravasz, D. C. V. Essen, K. Knoblauch, Z. Toroczkai, and H. Kennedy, "Cortical High-Density Counterstream Architectures," *Science*, vol. 342, no. 6158, p. 1238406, Nov. 2013.

[12] P. R. Montague, S. E. Hyman, and J. D. Cohen, "Computational roles for dopamine in behavioural control," *Nature*, vol. 431, no. 7010, pp. 760–767, Oct. 2004.

[13] W. V. Quine, *From a Logical Point of View: Nine Logico-Philosophical Essays, Second Revised Edition*, Revised edition. Cambridge, Mass: Harvard University Press, 1980.

[14] M. Ramscar, M. Dye, and J. Klein, "Children Value Informativity Over Logic in Word Learning," *Psychological Science*, p. 0956797612460691, Apr. 2013.

[15] W. Nichols, "Advertising Analytics 2.0 - Harvard Business Review," *Harvard Business Review*, 2013. [Online]. Available: http://hbr.org/2013/03/advertising-analytics-20/. [Accessed: 09-May-2013].

[16] D. Kahneman, P. P. Wakker, and R. Sarin, "Back to Bentham? Explorations of Experienced Utility," *The Quarterly Journal of Economics*, vol. 112, no. 2, pp. 375–406, May 1997.

[17] J. Suomala, L. Palokangas, S. Leminen, M. Westerlund, J. Heinonen, and J. Numminen, "Neuromarketing: Understanding Customers Subconscious Responses to Marketing," *Technology Innovation Management Review*, December , pp. 12–21, 2012.