Available online at www.sciencedirect.com
ScienceDirect
Neural Networks 20 (2007) 668–675
Neural Networks
www.elsevier.com/locate/neunet

Multiple model-based reinforcement learning explains dopamine neuronal activity

Mathieu Bertin[a,b,*], Nicolas Schweighofer[c], Kenji Doya[a,d]

[a] *ATR Computational Neuroscience Labs, 2-2-2 Hikaridai, "Keihanna Science City", Kyoto 619-0288, Japan*
[b] *Laboratoire d'Informatique de Paris 6, Universite Paris 6 Pierre et Marie Curie, 4 place Jussieu 75005, Paris, France*
[c] *Department of Biokinesiology and Physical Therapy, University of Southern California, 1540 E. Alcazar St. CHP 155, Los Angeles 90089-9006, USA*
[d] *Neural Computation Unit, Initial Research Project Laboratory, Okinawa Institute of Science and Technology, 12-22 Suzaki, Gushikawa, Okinawa, 904-2234, Japan*

Received 18 February 2005; accepted 11 April 2007

Abstract

A number of computational models have explained the behavior of dopamine neurons in terms of temporal difference learning. However, earlier models cannot account for recent results of conditioning experiments; specifically, the behavior of dopamine neurons in case of variation of the interval between a cue stimulus and a reward has not been satisfyingly accounted for. We address this problem by using a modular architecture, in which each module consists of a reward predictor and a value estimator. A "responsibility signal", computed from the accuracy of the predictions of the reward predictors, is used to weight the contributions and learning of the value estimators. This multiple-model architecture gives an accurate account of the behavior of dopamine neurons in two specific experiments: when the reward is delivered earlier than expected, and when the stimulus–reward interval varies uniformly over a fixed range.

Keywords: Dopamine; Reinforcement learning; Multiple model; Timing prediction; Classical conditioning

1. Introduction

Reacting correctly to its environment requires an animal to continuously anticipate the consequences of its observations and actions. Understanding how these predictions are constructed, through statistical inference of the current observations and memory of past experiences, is therefore critical. In the simple case of classical conditioning experiments, a reward is delivered shortly after a cue stimulus. Through repeated pairing of this conditioned stimulus and a reward, the animal learns to use the stimulus as a predictor of the occurrence and timing of the following reward. The issue we address in this paper is the prediction of the precise timing of the stimulus–reward interval (SRI). We propose a new model showing how animals can learn to associate a number of possible SRI to a single stimulus. Our model notably offers an explanation for two otherwise puzzling experimental results concerning the role of dopamine when the SRI varies.

Direct evidence from electrophysiological recordings in monkeys (Montague, Dayan, & Sejnowski, 1996; Schultz, 1998) and indirect evidence from fMRI studies in humans (Pagnoni, Zink, Montague, & Berns, 2002) during these simple conditioning experiments strongly suggest that the activity of dopamine (DA) neurons encode the error between predicted reward and actual reward. Early in training, a burst of DA neurons activity occurs at the time of the reward delivery. As training progresses, this burst disappears, and instead a burst of activity occurs at the time of the cue stimulus. If however, the reward is unexpectedly not delivered in one trial, there is a "dip" in DA activity, precisely at the time when the reward was supposed to be delivered.

Several early computational models (Moore et al., 1986; Sutton & Barto, 1990) use temporal difference (TD) methods (Sutton, 1988) to describe experimental results of the conditioned nictitating membrane response. Application of TD learning theory to DA measurements in later studies

* Corresponding author at: ATR Computational Neuroscience Labs, 2-2-2 Hikaridai, "Keihanna Science City", Kyoto 619-0288, Japan. Tel.: +81 774 95 1235; fax: +81 774 95 1259.
E-mail address: mbertin@atr.jp (M. Bertin).

0893-6080/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2007.04.028

(Daw & Touretzky, 2002; Montague et al., 1996; Schultz, Dayan, & Montague, 1997; Suri & Schultz, 1999), could accurately reproduce DA neuron activity during simple conditioning in terms of prediction error. TD learning is a real-time learning strategy aiming at building accurate predictions based on past experience. The predictions are computed as a "value" function, a sum of the expected future rewards. At each instant, predictions are compared to actual outcomes; the error in prediction (TD error) is then used to update the value function.

In these earlier implementations of the TD learning theory – for consistency, we will only refer to the tapped delay line model (Montague et al., 1996) – time is sequenced in steps. The current state is implemented as a row vector $s(t)$ with $s_i(t) = 1$ if i is the time steps elapsed since the stimulus, and $s_j(t) = 0$ otherwise. At each time step, the agent builds a value function, $V(t)$, prediction of future (discounted) rewards:

$$V(t_0) = \sum_{t=t_0}^{\infty} \gamma^t r(t) \tag{1}$$

where γ $(0 < \gamma < 1)$ is a discounting parameter. Note that in the typical simple conditioning experiment, there is only one reward per trial, and the value function simply equals one discounted reward.

In neural network implementations, the value function of the current state is computed by the inner product:

$$V(t) = s(t) * (w(t))^{\mathrm{T}} \tag{2}$$

with $w(t)$ a weight row vector, and $(w(t))^{\mathrm{T}}$ its transpose. Through learning, these weights are updated at each time step according to the current prediction error:

$$w(t) = w(t) + \eta s(t)\delta(t) \tag{3}$$

where η a learning rate and $\delta(t)$ the TD error (scalar), which models the DA neurons' activity, and is given by:

$$\delta(t) = r(t) + \gamma V(t+1) - V(t). \tag{4}$$

During learning, the agent gradually builds a value function that correctly predicts the incoming reward. After learning, if the reward is given, the TD error is null at all time, except at the time of the conditioned stimulus. If however a reward is not given, the TD error is negative at the time the reward was expected.

Thus, the TD error given by these earlier models reproduces the DA neurons' activity remarkably well in the simple conditioning experiments. These models however fail to account for two recent experimental results in which the intervals between the conditioning stimulus and the reward are varied. We now describe these two experimental conditions on temporal variability: (1) earlier reward delivery, and (2) uniform variation of the stimulus–reward interval.

2. Experiments on temporal variability

2.1. Earlier reward delivery

In experiments conducted on dopamine measurements (Hollerman & Schultz, 1998), a monkey is trained to expect a reward precisely one second after the conditioned stimulus. After training, the reward is suddenly presented 0.5 s early or late. Three different types of DA responses were found (see Fig. 1).

- if the reward is given when expected, no change in DA activity is visible.
- if the reward is given late, a "dip" in DA activity occurs at the time the reward was expected; then there is a burst of activity shortly after the reward is finally given.
- if the reward is given early, a burst marks the reward delivery; however, no significant dip is observed at the time the reward was expected.

These observations follow the previous conclusions on the experiment (Hollerman & Schultz, 1998). While we will adhere to the authors' claims, we would state a few precautions concerning the use of this figure. Some of the results are undisputable, such as the presence of a dopamine burst shortly after early and delayed rewards. The dip of activity in the case of a late reward is also described as "significant" by the initial authors. Other results, however, are suject to qualitative interpretation. For example, we could not quantitatively detemine if one of the two bursts (early or late) is higher than the other. More importantly, the authors state that no dip of activity is observed following earlier reward. Although we will adhere to this observation in the current paper, further experiment might be needed to ascertain this result in a more quantitative way.

The tapped delay line model (Montague et al., 1996) reproduces accurately the first two types of responses, but fails in the last: it predicts a pause in DA activity at the time when the reward is usually expected. This happens because the agent only reacts time-step by time-step, and thus cannot infer that the reward it received earlier is the one it was expecting. In order to explain this experimental data, a new model based on a semi-Markov architecture has recently been proposed (Courville, Daw, & Touretzky, 2004). In this model, only two states are considered (ISI and ITI, inter stimulus and inter trial interval), and the agent tries through learning to predict the duration and probability of these two states. This model reproduces the above data accurately, because once the reward has been given, the agent does not expect any more rewards—thus, in the third condition, no dip of activity is created at the time the reward was given during training.

2.2. Uniformly varying stimulus–reward interval

In another experiment on DA measurement on monkeys (Fiorillo & Schultz, 2001), the interval between the stimulus and the reward varies uniformly over a fixed range (1–3 s) throughout training. Fig. 2 shows the response of a DA neuron during this experiment. When the SRI is short (lower part of the figure), a strong burst of activity marks the time of reward; for longer SRI (higher part of the figure), the observed burst of activity is lower. For the longest intervals, it appears impossible to state if there is actually a positive response.

Earlier TD models do not account for these results. In the Montague model, for instance, the value function is reorganized