Higher-Order Predictive Information for Learning an Infinite Stream of Episodes

Byoung-Tak Zhang
School of Computer Science and Engineering &
Cognitive Science and Brain Science Programs
Seoul National University
Seoul, 151-742, Korea
btzhang@bi.snu.ac.kr

Extended Abstract

We consider the problem of lifelong learning from an indefinite stream of temporal episodes, i.e. a time series consisting of episodes, where the number of the episodes is potentially infinite and the length of each episode varies. Examples of this class of learning include a humanoid robot that continually learns to imitate various human behaviors [7], a computer music system that learns to compose from a continuous stream of music pieces [4], and a cognitive system that incrementally learns visual concepts from a series of movies over a long period of time [9].

What kinds of objective function should the lifelong learner use to balance the short-term and long-term performance? How should the learner optimize its model complexity when the statistics of the episodes change over time? Maximization of the expected future reward, such as a value function used in reinforcement learning, might be useful if we could define rewards for a prespecified goal. For learning an indefinite stream of episodes, we find the mutual information-based measures of information theory, such as predictive information [2], $I(X_{future}; X_{past})$, and empowerment [3], $\max_A I(X_{t+1}; A_t|x_t)$, suitable. The predictive information is, however, typically approximated by restricting the time horizons to a single time step, i.e. $I(X_{future}; X_{past}) = I(X_{t+1}; X_t)$. Though this is exact under the Markov assumption, i.e. the probability of a state depends only on the probability of the previous state, and still can generate explorative behavior [1], the predictive power can be improved by increasing the order of temporal dependency.

Here we extend the first-order predictive information to the *k*th-order predictive information for lifelong learning from a continuous stream of time-series episodes. We generalize the predictive information $I(X_{t+1}; X_t)$ by replacing the first-order term X_t for past by *k*-th order history term $X_{t-k:t} = X_{t-k,t-k+1,\ldots,t}$, i.e. $I(X_{t+1}; X_{t-k:t})$. The generalized, higher-order predictive information is written as:

$$\begin{aligned} I(X_{t+1}; X_{t-k:t}) &= \langle \log_2 \frac{P(x_{t+1}, x_{t-k:t})}{P(x_{t+1})P(x_{t-k:t})} \rangle \\ &= \langle \log_2 \frac{P(x_{t+1}|x_{t-k:t})}{P(x_{t+1})} \rangle \\ &= \iint P(x_{t+1}, x_{t-k:t}) \log_2 \frac{P(x_{t+1}|x_{t-k:t})}{P(x_{t+1})} dx_{t+1} dx_{t-k:t} \end{aligned}$$

Specifically, we consider the problem of learning human-like behavior of a humanoid robot (Darwin OP) consisting of 20 joint angles of arms and legs. The goal of the robot is to master, over an extended period of time, a large repertoire of flexible behaviors, such as walking, running, kicking, expressing greetings, saying good-bye, hand-shaking, etc. A higher-order Markov model is trained on a sequence of trajectories of robot joint angles using the

higher-order predictive information as the objective function. We compare the higher-order predictive models with the standard (first-order) predictive model, showing that the higher-order models can enhance the predictive power. We analyze the evolution of the predictive information as a function of the number of learned episodes to characterize the inherent complexity of the learning task [2].

One problem with the higher-order predictive information is that it is hard to compute since the number of parameters to estimate grows exponentially with the order. However, recent findings in machine learning with higher-order models suggest that, for the purpose of learning, the *k*th-order Markov models with a small *k*, e.g. $k << n$, offer an effective model that can exploit the additional history without facing the combinatorial problem for large *k*th-order Markov models [8]. Harnessing these findings, we use a hypergraph model that incrementally approximates the higher-order Markov model from a stream of multidimensional temporal episodes of robot-behavior trajectories using an importance sampling-based Monte Carlo approximation or a particle filter. This allows us to estimate the higher-order predictive information as follows:

$$
\begin{aligned}
I(X_{t+1}; X_{t-k:t}) &= \iint P(x_{t+1}, x_{t-k:t}) \log_2 \frac{P(x_{t+1}|x_{t-k:t})}{P(x_{t+1})} dx_{t+1} dx_{t-k:t} \\
&= \iint \frac{P(x_{t+1}, x_{t-k:t})}{Q(x_{t+1}, x_{t-k:t})} Q(x_{t+1}, x_{t-k:t}) \log_2 \frac{P(x_{t+1}|x_{t-k:t})}{P(x_{t+1})} dx_{t+1} dx_{t-k:t} \\
&\approx \sum_{i=1}^{N} w^{(i)} \log_2 \frac{P(x_{t+1}^{(i)}|x_{t-k:t}^{(i)})}{P(x_{t+1}^{(i)})}
\end{aligned}
$$

where $Q(x_{t+1}, x_{t-k:t})$is the approximating distribution from which the particles (samples) are drawn, i.e. $x^{(i)} = (x_{t+1}^{(i)}, x_{t-k:t}^{(i)}) \sim Q(x_{t+1}, x_{t-k:t})$, N is the number of particles, $w^{(i)} = P(x_{t+1}^{(i)}, x_{t-k:t}^{(i)}) / Q(x_{t+1}^{(i)}, x_{t-k:t}^{(i)})$ is the importance weight of particle i.

We also discuss the potential application of the higher-order extension of the predictive information to other information-theoretic approaches to lifelong learning with the perception-action cycle, (X_t, A_t), t = 0, 1, …, especially those that extend the conventional reinforcement learning framework by predictive information or empowerment based on the free-energy principle, i.e. $F(X_t, A_t, \beta) = I(X_t; A_t) - \beta Q(X_t, A_t)$, as explored in some previous work [6, 5].

Acknowledgements: This work was supported in part by the National Research Foundation (NRF-2010-0017734) and the AFOSR/AOARD R&D Grant 124087.

References

[1] Ay, N., Bertschinger, N., Der, R., Guetter, F., & Olbrich, E., Predictive information and explorative behavior in autonomous robots, *European Physical Journal* B, 63:329--339, 2008.

[2] Bialek, W., Nemenman, I., & Tishby, N., Predictability, complexity, and learning, *Neural Computation*, 13:2409-2463, 2001.

[3] Jung, T., Polani, D. & Stone, P., Empowerment for continuous agent-environment systems, *Adaptive Behavior*, 19(1):16-39, 2011.

[4] Rebuschat, P., Rohrmeier, M., Hawkins, J. A., & Cross, I. (Eds.), *Language and Music as Cognitive Systems*, Oxford Uniersity Press, 2011.

[5] Still, S. & Precup, D.. An Information-theoretic approach to curiosity-driven reinforcement learning, *Theory in Biosciences*, 131(3):139-148, 2012.

[6] Tishby, N. & Polani, D., Information theory of decisions and actions. In: *Perception-Reason-Action Cycle: Models, Algorithm and Systems*. Springer, 2010.

[7] Yi, S.-J., McGill, S., Zhang, B.-T., Hong, D., & Lee, D. D., Active stabilization of a humanoid robot for real-time imitation of a human operator, In *Proceedings of 12th IEEE-RAS International Conference on Humanoid Robots*, 2012.

[8] Zhang, B.-T., Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory, *IEEE Computational Intelligence Magazine*, 3(3):49-63, 2008.

[9] Zhang, B.-T., Ha, J.-W., & Kang, M., Sparse population code models of word learning in concept drift, In *Proc. 34th Annual Conference of the Cognitive Science Society (CogSci 2012)*, pp. 1221-1226, 2012.