Expert Systems with Applications 41 (2014) 5212–5226

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A Dynamic-Bayesian Network framework for modeling and evaluating learning from observation

Santiago Ontañón [a,*], José L. Montaña [b], Avelino J. Gonzalez [c]

[a] Drexel University, Philadelphia, PA, USA
[b] University of Cantabria, Santander, Spain
[c] University of Central Florida, Orlando, FL, USA

ARTICLE INFO

Keywords:
Learning from observation
Dynamic Bayesian Networks

ABSTRACT

Learning from observation (LfO), also known as *learning from demonstration*, studies how computers can learn to perform complex tasks by observing and thereafter imitating the performance of a human actor. Although there has been a significant amount of research in this area, there is no agreement on a unified terminology or evaluation procedure. In this paper, we present a theoretical framework based on Dynamic-Bayesian Networks (DBNs) for the quantitative modeling and evaluation of LfO tasks. Additionally, we provide evidence showing that: (1) the information captured through the observation of agent behaviors occurs as the realization of a stochastic process (and often not just as a sample of a state-to-action map); (2) learning can be simplified by introducing dynamic Bayesian models with hidden states for which the learning and model evaluation tasks can be reduced to minimization and estimation of some stochastic similarity measures such as crossed entropy.

1. Introduction

Learning by watching others do something is a natural and highly effective way for humans to learn. It is also an intuitive and highly promising avenue for machine learning. It provides a way for machines to learn how to perform tasks in a more natural fashion. For many tasks, learning from observation is more natural than providing static examples that explicitly contain the solution, as in the traditional supervised learning approach. It is also easier than manually creating a controller that encodes the desired behavior. Humans typically just perform the task and trust that the observer can figure out how to successfully imitate the behavior.

Although there has been a significant amount of research in learning from observation (LfO), there is no agreement on a unified terminology. Works reported in the literature also refer to *learning from demonstration*, *learning by imitation*, *programming by demonstration*, or *apprenticeship learning*, as largely synonymous to learning from observation. In learning from demonstration, a human purposely demonstrates how to perform a task or an action, expressly to teach a computer agent how to perform the same task or mission. We consider learning from demonstration to be a specialization of LfO and define the latter as a more general learning approach, where the actor being observed need not be a willing participant in the teaching process.

Specifically, the problem we are trying to address in this paper is the lack of a unified framework to understand existing work in LfO, as well as the lack of standard evaluation metrics to assess the performance of LfO algorithms (which are typically evaluated using metrics designed for standard supervised learning). To that purpose, we present an unified framework for learning from observation based on *Dynamic Bayesian Networks* (DBNs) (Nefian, Liang, Pi, Liu, & Murphy, 2002). We provide both an intuitive description of the framework, as well as a formal statistical model of LfO. The main contributions of this paper, are:

- A formal statistical model of LfO, that provides a unified vocabulary and theoretical framework for LfO.
- A taxonomy of the different behaviors that can be learned through LfO.
- An explicit formulation of the difference between supervised learning and LfO algorithms. This is important because in most LfO work, standard supervised algorithms (like neural networks, or nearest-neighbor classifiers) are used, yet there are many behaviors to be learned through LFO for which those algorithms are not appropriate.

* Corresponding author. Tel.: +1 2155714109.
E-mail addresses: santi@cs.drexel.edu (S. Ontañón), montanjl@unican.es (J.L. Montaña), gonzalez@ucf.edu (A.J. Gonzalez).

http://dx.doi.org/10.1016/j.eswa.2014.02.049
0957-4174/

- A proposal for standard evaluation metrics for agents trained through LfO (currently lacking from the literature). Our framework makes explicit the reason for which standard metrics, such as classification accuracy, do not properly reflect how well an LfO algorithms can learn complex tasks in some situations. We describe the reasons for this, and propose an alternative evaluation approach based on the Kullback-Liebler divergence.

The remainder of this paper is organized as follows. Section 2 briefly summarizes previous research in the field. After that, Section 3 introduces a common framework and vocabulary for learning from observation, including a statistical formalization of the problem. Section 4 focuses on evaluation metrics for LfO algorithms. Finally, Section 5 presents an empirical validation of two of our claims: (a) supervised learning algorithms are not appropriate for some LfO behaviors, and (b) our proposed evaluation metric is more accurate than the typical metrics used in the literature of LfO.

2. Learning from observation background

LfO is a subfield of machine learning that studies how to learn behavior from observation or demonstration. This has many practical applications. A first set of interesting applications of LfO concern allowing machines to learn how to perform complex behaviors that would be difficult to manually program (e.g. driving vehicles Pomerleau, 1989, robotics Argall, Chernova, Veloso, & Browning, 2009, playing videogames Ontañón, Mishra, Sugandh, & Ram, 2010). Another interesting set of applications involves understanding and analyzing behavior, i.e., using LfO to generate models of certain behaviors of interest in order to compare them, cluster them or understand them (Dereszynski, Fern, Dietterich, Hoang, & Udarbe, 2011).

In this paper, we argue that LfO is fundamentally different from traditional supervised, unsupervised and reinforcement learning approaches. Clearly, LfO differs from unsupervised learning because the examples (demonstrations) contain an implicit indication of correct behavior. Furthermore, LfO clearly differs from reinforcement learning because LfO learns from a collection of traces (or trajectories), rather than from trial and error through a reinforcement signal (although, as we show later, some reinforcement learning approaches use LfO as a way to guide the learning process, e.g., Lin, 1992).

LfO can be more readily likened to supervised learning, but two key differences exist. First, in LfO the learning examples are time-based, continuous and not easily separable throughout the duration of the exercise. Furthermore, no explicit linkage between cause and effect is provided, but must be extracted automatically by the learning algorithm. The cause of an action might very well be something perceived in a past instant of time, rather than in the current perceptual state. Thus, the form of supervised learning that is more related to LfO is that of *sequential learning* (Dietterich, 2002).

A second key difference is on what needs to be learned. Supervised machine learning techniques (including sequential ones) focus on learning models that minimize the prediction error, i.e., they learn to predict the output of the learning task given the input, on average. However, that is not the goal in LfO. Consider an agent trying to learn by observing the behavior of an actor who, while driving a car, chooses a different random speed each minute with a mean of 100kph and a certain variance. A standard supervised learning method would learn to predict that the speed should always be 100kph (because that is the value that yields minimum prediction error). Therefore, an LfO agent should learn that the speed must be changed when appropriate with a particular variance. LfO aims to replicate the behavior of the actor, rather than to minimize the prediction error. For that reason, supervised learning techniques for LfO have to be employed with care, and a different class of algorithms is required in the general case. However, as our previous work shows, some types of learning from observation tasks can be addressed with supervised learning (for example, as we explain later, when learning state-less deterministic behavior, minimizing prediction error is equivalent to replicating behavior).

Work in learning from observation can be traced back to the early days of AI. Bauer (1979) proposed in 1979 to learn programs from example executions, which basically amounts to learning strategies to perform abstract computations by demonstration. This form of learning has been especially popular in robotics (Lozano-Pérez, 1983). Other early mentions of learning from observation come from Michalski and Stepp (1983) who define it merely as unsupervised learning, and from Pomerleau (1989), who developed the ALVINN system that trained neural networks from observation of a road-following automobile in the real world.

More recent work on the more general LfO subject came from Sammut, Hurst, Kedzier, and Michie (1992), Sidani (1994) and Gonzalez, Georgiopoulos, DeMara, Henninger, and Gerber (1998). Fernlund, Gonzalez, Georgiopoulos, and DeMara (2006) used learning from observation to build agents capable of driving a simulated automobile in a city environment. The neural network approach to learning from observation has remained popular, and contributions are still being made, such as the work of Moriarty and Gonzalez (2009), in the context of computer games.

In robotics, learning from demonstration has been extensively used to implement human behavior in humanoid robot movements. Bentivegna and Atkeson (2001) used learning from demonstration to teach a humanoid robot to play air hockey using a set of action primitives, each describing a certain basic action. Chernova and Veloso (2007, 2008) studied the problem of multi-robot learning from demonstration where a group of agents was shown how to collaborate with each other. Other important work was reported by Schaal (1996), Atkeson et al. (2000) and Argall et al. (2009) among many others.

Könik and Laird (2006) studied learning from observation with the SOAR system by using inductive logic programming techniques.

A theoretical approach to LfO is that of Khardon (1999), where he proposed to use a systematic algorithm that enumerates all the possible finite-state machines (given the inputs and outputs of a given domain), and rank them according to their probability of achieving the goal, as well as to how consistent they are with the behavior observed from the expert. An important difference between Khardon's work and most work on LfO is that Khardon assumed that the learning agent has access to a description of the goal (i.e., the learning agent knows, during learning, if a particular behavior would achieve the goal or not in a given scenario). Most work on LfO assumes that the only form of input are samples of behavior from the expert, without any explicit description of the goal to be achieved.

Other significant work done under the label of learning from demonstration has emerged recently in the case-based reasoning (CBR) community. Floyd, Esfandiari, and Lam (2008) presented an approach to learn how to play RoboSoccer by observing the play of other teams. Ontañón and associates (Ontañón, Mishra, Sugandh, & Ram, 2007, 2010, 2012) used learning from demonstration in the context of case-based planning, applied to real-time strategy games. Rubin and Watson (2011) used LfO for creating a Poker-playing agent. And Lamontagne, Rugamba, and Mineau (2012) studied techniques based on conditional entropy to improve case acquisition in CBR-based LfO. The main difference between the work based on CBR and the previous work presented in this section is that CBR methods are related to lazy machine learning techniques that do not require any form of generalization