Expert Systems With Applications 112 (2018) 331–341

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Combination of learning from non-optimal demonstrations and feedbacks using inverse reinforcement learning and Bayesian policy improvement

Ali Ezzeddine[a,*], Nafee Mourad[b], Babak Nadjar Araabi[a,*], Majid Nili Ahmadabadi[b]

[a] Machine Learning and Computational Modeling Laboratory, School of ECE, College of Engineering, University of Tehran, Tehran, Iran
[b] Cognitive Systems Laboratory, School of ECE, College of Engineering, University of Tehran, Tehran, Iran

ARTICLE INFO

Article history:
Received 17 March 2018
Revised 24 May 2018
Accepted 14 June 2018
Available online 23 June 2018

Keywords:
Teaching by demonstrations
Inverse reinforcement learning
Interactive learning
Human evaluative feedbacks

ABSTRACT

Inverse reinforcement learning (*IRL*) is a powerful tool for teaching by demonstrations, provided that sufficiently diverse and optimal demonstrations are given, and learner agent correctly perceives those demonstrations. These conditions are hard to meet in practice; as a trainer cannot cover all possibilities by demonstrations, he may partially fail to follow the optimal behavior. Also, trainer and learner have different perceptions of the environment including trainer's actions. A practical way to overcome these problems is using a combination of trainer's demonstrations and feedbacks.

We propose an interactive learning approach to overcome the challenge of non-optimal demonstrations by integrating human evaluative feedbacks with the *IRL* process, given sufficiently diverse demonstrations and the domain transition model. To this end, we develop a probabilistic model of human feedbacks and iteratively improve the agent policy using Bayes rule. We then integrate this information in an extended *IRL* algorithm to enhance the learned reward function.

We examine the developed approach in one experimental and two simulated tasks; i.e., a grid world navigation, a highway car driving system and a navigation task by the e-puck robot. Obtained results show significant improved efficiency of the proposed approach in face of having different levels of non-optimality in demonstrations and the number of evaluative feedbacks.

1. Introduction

Machine learning aims at training complex systems, such as autonomous cars and assistant robots, to perform sophisticated tasks in the real world. In addition, performing such real world applications necessitates adaptation to user's preferences as well as to changes in tasks and in the environment, with minimum user intervention. Researches in teaching by demonstrations and interactive learning target that aim and these adaptation capabilities; as a powerful replacement for manual coding and behavior tuning.

Interactive learning techniques can be categorized into two major clusters; i.e., Learning from Feedbacks (*LfF*) and Learning from Demonstrations (*LfD*). In *LfF*, the trainer evaluates the learner behavior and gives feedbacks in different formats (e.g., binary reward, numeric reward, etc.) to improve the learner policy (i.e., state-action mapping). In *LfD*, the agent tries to learn its policy by observing the trainer demonstrations. These techniques facilitate improving artificial systems' behavior by non-technical users.

Simplicity of providing evaluative feedbacks and absence of correspondence problem (Nehaniv & Dautenhahn, 2007) between the trainer and the learner are the main advantages of *LfF*. Nevertheless, *LfF*, in its basic form, suffers from the curse of dimensionality and the learner's random behavior at early learning trials; when learning starts from scratch. It means that, *LfF* requires a large number of evaluative feedbacks and has a slow convergence rate to the desired policy.

In *LfD* technique, in contrast, the correct action can be directly provided by the demonstrator; this reduces the learner exploration in those states it receives correct demonstrations. This property results in faster learning provided that, demonstrations are correct and can be generalized to all possible situations. *LfD* techniques handle the transfer of new behaviors from trainers to learning agents, and unlike the record and replay methods, *LfD* techniques generalize from local state-actions to the whole space. Nevertheless, the correspondence problem should be resolved prior to using *LfD* methods.

* Corresponding authors.
E-mail addresses: a.ezedin@ut.ac.ir (A. Ezzeddine), n.mourad@ut.ac.ir (N. Mourad), araabi@ut.ac.ir (B.N. Araabi), mnili@ut.ac.ir (M.N. Ahmadabadi).

https://doi.org/10.1016/j.eswa.2018.06.035
0957-4174/

LfD techniques can be classified into two main categories according to the mechanism employed to imitate the demonstrator: "Direct imitation learning" and "Apprenticeship learning" methods. "Direct imitation learning" comprises those methods that use supervised learning algorithms to directly derive the policy; see (Billard, Calinon, Dillmann, & Schaal, 2008; Hajimirsadeghi, Ahmadabadi, Araabi, & Moradi, 2012; M Ali beigi, 2015). On the other hand, "Apprenticeship learning" methods (Abbeel & Ng, 2004) (also called intention learning or indirect imitation learning) are framed as Inverse Reinforcement Learning (*IRL*) problems (Ng & Russell, 2000) and aim to generalize the observed demonstrations by estimating the reward function that shapes the same behavior. Thereafter, the policy that maximizes the expected sum of this reward is derived by planning algorithms, such as dynamic programming (*DP*) (Sutton & Barto, 1998). Estimation of the reward function, to some extent, brings in robustness against changes in the learner configuration and the environment.

Performance of *IRL* techniques mainly depends on the quality of demonstrations. Here, quality refers to sufficiency and optimality of demonstrations and is defined from the learner perspective. A demonstration is called optimal if it is generated by following the optimal policy. A set of demonstrations is sufficient if *IRL* output is sufficiently generalized to all states.

In practice, demonstrations can be non-optimal due to miscellaneous reasons: (a) difficulty in demonstrating some actions by the human trainer due to the complexity of the task. (b) Expertise requirement in the relevant field in order to guide the agent. (c) Correspondence problem due to some differences in physical embodiment and perception of the trainer and the learner. (d) Infeasibility to provide some demonstrations due to dangerous and harmful conditions for the human trainer. (e) Noisy demonstrations due to the environmental disturbances and (f) the learner's imperfect perception. It means, the learner should deal with non-optimal demonstrations in the real world.

There are different *IRL* methods that perform well in training the agent in parts of the state space where optimal, or close to optimal demonstrations are given. However, assuming to have perfect demonstrations is not realistic in practice; therefore, additional information is needed to compensate for non-optimal demonstrations. Using human feedbacks is our choice to overcome the challenges of non-optimality. By doing so, we benefit from complementary properties of *LfF* and *LfD* methods in realistic situations.

In this work, we assume that the agent learns from non-optimal demonstrations through *IRL* and improves its performance further by receiving human evaluative feedbacks; i.e., right/wrong instructions. The question here is how to integrate human feedbacks with *IRL* in order to reduce the sensitivity to inaccurate demonstrations. The answer to this question is the main contribution of this paper.

The rest of the paper is structured as follows: Section 2 deals with the review of some related works. Section 3 gives the necessary definitions and assumptions associated with our work. Section 4 gives details of our proposed approach. Section 5 provides the evaluation metrics for testing the performance of the proposed method and shows the experimental results in some benchmarks. And finally the paper is concluded in Section 6.

2. Related works

2.1. Learning from demonstrations (IRL approaches)

All of the *IRL* approaches have the same principle of estimating a reward function that motivates the demonstrated behavior, but differ in their methodology to do so. Some works in the *IRL* focus on learning the reward function under which the expected feature count of demonstrations and the learned policy are matched (Abbeel & Ng, 2004; Ratliff, Bagnell, & Zinkevich, 2006; Syed, Bowling, & Schapire, 2008). Some others introduce a probability model over demonstrations and seek to maximize the probability of the reward under this model (Babes, Marivate, Subramanian, & Littman, 2011; Boularias, Kober & Peters, 2011; Ramachandran & Amir, 2007; Ziebart, Maas, Bagnell, & Dey, 2008). In addition, there exists other approaches that mainly focus on solving the challenges of *IRL* algorithms, like the large computational complexity (Sharma, Kitani, & Groeger, 2017), scaling to a large environment space (Finn, Levine, & Abbeel, 2016), solving the *IRL* without predefining the space's features (Wulfmeier, Ondruska, & Posner, 2015), etc. However, almost all of these foresaid approaches use the traditional assumption of the *IRL* that demonstrations are optimal or near to the optimal; i.e., they do not take into account the presence of non-optimality in demonstrations. Since usually providing optimal demonstrations is not possible in practice, in our work, we mainly focus on overcoming the non-optimality challenge in the demonstrations so as to enhance learning.

On the other hand, few *IRL* works have incorporated the assumption of non-optimality in their learning process. In (Silver, Bagnell, & Stentz, 2010), the problem of non-optimality is handled by relaxing the constraints of the object function with the assumption that the trainer demonstrations define a corridor in which the optimal path exists. Using this approach, the non-optimality in demonstrations can be overcome provided that its level is at a small scale. (Zheng, Liu, & Ni, 2014) deals with demonstrations that are adversely affected by a sparse noise and considers non-optimality in only some demonstrations. This work proposes a model to identify and differentiate between noisy and reliable demonstrations. Unlike these approaches, in our method, we assume that the non-optimality exists in each demonstration and it is impossible to differentiate between optimal and non-optimal ones. Also, we consider its level to be more than noise. In (Xia & El Kamel, 2016), non-optimal demonstrations are pretreated and improved by means of applying maximum a posteriori, and are later used in the *IRL* algorithm. However, in this approach, for non-optimalities more than noise level, pretreating demonstrations is not effective. In (Coates, Abbeel, & Ng, 2008), it is suggested that a large number of non-optimal demonstrations could implicitly encode the optimal behavior and accordingly a generative model is used to derive optimal demonstrations from a given set. Here, it is not evident that how much the provided demonstrations can be distant from the optimal ones. Almost all of these works, in addition to some Bayesian *IRL* approaches, assume that the level of non-optimality is small, a large number of demonstrations are available, and/or non-optimality exists in some demonstrations. These algorithms can only withstand a margin of non-optimality and they just deal with noisy demonstrations.

In this work, our assumption of non-optimality is extended to include more than noisy demonstrations, in the sense that other non-optimality factors (as previously mentioned in introduction) are also considered in the learning process as well. In addition, the non-optimality can exist in each demonstration. Accordingly, this assumption increases the level of non-optimality which can't be improved with additional or a large number of demonstrations and can't be overcome unless another source of information is added to the learning process. Therefore, our approach attempts to extend the *IRL* by benefiting from another source of information based on human feedbacks in order to improve the learning in face of non-optimality in demonstrations.

2.2. Learning from human feedbacks

Regarding the use of human feedbacks in learning a policy, here we review some of the works that treat human feedbacks as either numerical rewards (values) or evaluative feedbacks (right/wrong). From the works that deal with numerical rewards as their only