Learning from Disagreeing Demonstrators

Bruno N. da Silva
Computer Science Department
University of British Columbia
2366 Main Mall, Vancouver, BC, Canada
bnds@cs.ubc.ca

Abstract

We study the problem of learning from disagreeing demonstrators. We present a model that suggests how it might be possible to design an incentive-compatible mechanism that combines demonstrations from human agents who disagree on the evaluation of the demonstrated task. Apart from comonotonicity of preferences over atomic outcomes, we make no assumptions over the preferences of our demonstrators. We then suggest that a reputation mechanism is sufficient to elicit cooperative behavior from otherwise competitive human agents.

Introduction

Task demonstration is a promising approach for dealing with the difficulty of robot programming in complex settings. Instead of placing the integrity of the process' burden on the learning robot, a fraction of it is assigned to humans by the teaching responsibility. This approach is inspired by the mimicking behavior witnessed in nature and takes advantages of the human cultural expertise in transmitting knowledge though demonstration (Breazeal and Scassellati, 2007).

Just like in the literature of supervised learning, the demonstration mechanism allows a set of examples to be used by robots to learn and produce a general policy. However, one of the differences in teaching robots by demonstration is that it offers an opportunity for humans to criticize the policies generated by the robot (Argall, Browning and Veloso, 2007). This 'contextual criticism' seems to increase the efficiency of the process, making the demonstration approach very appealing.

Unfortunately, a characterizing trait of human nature is the idiosyncrasies that distinguish each individual. For almost every task, we find ourselves with very particular perspectives, usually diverging about the desirability and preferences over different states, and sometimes even disagreeing about what would be the best strategy to reach a certain state. Therefore, we argue that if robots are to become a significant part of the human routine, it will be essential for them to deal with human peculiarities.

Motivated by this remark, this paper introduces a model where robots learn from human demonstrators who do not share a common preference over states of the world. As an inspiring example, imagine a married couple who tries to teach a robot how to drive their kids to school. This is a task that contains a number of traditional challenges usual in the Multiagent Systems community (including the problem of imperfect perception when identifying the correct state of the world, and the computation of which action to perform given an inferred decision point). However, we are mainly interested in a different aspect of this scenario: we assume that each of our human agents has a subjective policy for the task and they agree to disagree on the best strategy to transmit to the robot. More specifically, one of the human agents has a very aggressive driving style, while the other is too passive.

Clearly, no individual driving profile can be singled out a priori as better than the other. While in some cases a passive approach will diminish the risk of exposing the passengers to accidents, there may be situations where there is room for a more aggressive (i.e. less defensive) course of action that won't increase the likelihood of accidents by much, while incurring in a considerable decrease in the duration of the ride.

A straightforward way to solve this problem would be to give up efficiency and arbitrarily select one of the existing human agents to instruct the robot by demonstration. However, this would represent an unfair resolution of the problem since, in principle, no qualitative order exists over humans. Furthermore, we believe that since humans will delegate to the robot a task that is currently performed by them, a minimal trace of each demonstrator's driving style should be reflected in the robot behavior.

With these observations in mind, we design a framework that intelligently integrates inputs from our disagreeing sources and combine them into a single policy. In order to avoid a greedy equilibrium where each demonstrator ignores prospective combinations of driving styles, we will use a similar framework to (Argall, Browning and Veloso, 2007) and consider a critiquing phase in our mechanism. In this step, we encourage each demonstrator to carefully evaluate a policy generated from the poll of demonstrations of human agents. And in order to achieve incentive-compatibility, we include a reputation mechanism to the mode, in order to collect constructive criticism on the evaluation phase.

Copyright © 2009, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

A Model of Learning from Disagreeing Demonstrators

Model Parameters

We consider a set of humans agents D who wish to demonstrate to a single robot how to perform the task of driving their kids to school. All of them have unknown utility functions which are computed over a set of known aspects of the world O. In our model, we assume that the preference relations of the demonstrators are comonotonic over aspects of the world, i.e.

$$\forall d_i,d_j \in D,\ \forall O_k,\ o_a,o_b \in O_k \qquad o_a \succsim_i o_b \rightarrow o_a \succsim_j o_b.$$

In other words, for each outcome, we assume that our demonstrators agree on a weak ordering of its domain. However, they need not agree on preferences over combinations of the outcomes. In our example, assume we have two outcomes in our driving model: O_c is the number of crashes before reaching the school, and O_d is the duration of the ride. Therefore, as long as demonstrators agree, e.g., that 1) the smaller the number of crashes the better, and also that 2) short rides are preferable over longer ones, this would satisfy comonotonicity.

As for the robot agent, we assume that the robot's perception is faulty, and it recognizes each state of the world as dictated by an unknown mapping $H : S \rightarrow P$ which transforms states of the world in S into observations in P. On top of that, the goal of the robot is to construct a control policy $\pi : A \times P \rightarrow A$, from observations into actions in A. Since our objective is to allow demonstrations from humans to robots, we further assume that the mapping H is such that it allows a successful policy acquisition through demonstrations of the humans.

Procedure for Knowledge Acquisition

Our framework is based on (Argall, Browning and Veloso, 2007). As in that work, we assume knowledge transmission is performed in a two-stage process. In the first stage, as depicted in Figure 1, each human agent $d_i \in D$ directly demonstrates the task to the robot by executing it a finite number of times. For each execution in d_i's demonstration, the robot collects a sequence of (p_m,a_n) points. This pair maps the robot's perception p_m to the action a_n, which the demonstrator d_i regards as the best response to the current world state.

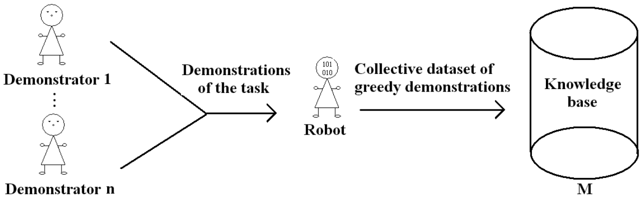
Demonstrator 1
⋮
Demonstrator n
Demonstrations of the task
101 010
Robot
Collective dataset of greedy demonstrations
Knowledge base
M

Figure 1. The first stage of our procedure, where a set of demonstrators execute the task to the robot, in order to allow the generation of a knowledge base M.

After this data collection, the robot possesses a set of demonstrations, each of which are in turn a collection of sequences of (p_m,a_n) pairs. From this set of demonstrations, it constructs a knowledge base M.

With the conclusion of this first stage, the robot possesses a set of action points in M which recommends candidate actions given state observations. For those perceived states which do not match any element in M, the robot can employ any heuristic similarity search procedure to infer an appropriate choice (e.g. a Nearest Neighbor search).

The problem with this set M is that it is a slack union of different perspectives of the task. This is because we assumed that each demonstrator executed the task without concerns about the behavior of fellow demonstrators. Therefore, a naïve policy based on this resulting set might display inconsistent actions due to random crossover combinations of very diverse behaviors.

This motivates the second stage of our procedure (Figure 2), which aims at polishing this set M into a new dataset that not only maintains a broad coverage of the demonstrators' perspective on the problem, but also has a more homogeneous behavior. In this stage, we introduce a critiquing step for the humans. Now, the robot is the one who simulates the series of executions of the task to the humans, and expects for each human an informative signal that indicates how they evaluate the most recent execution. Consequently, positive feedback from humans will strengthen the elements in M which contributed to the execution, while negative feedback weakens them. For this reason, each element in M is now coded as a (p_m,a_n,c) tuple, where c is a quantitative measure of the robot's confidence that a_n is a good response to p_m. Since the confidence of the (p_m,a_n) is affected by the critiques of all the demonstrators, the values in M after this stage will reflect a unified understanding of the task, as opposed to the segregated state of M before this critiquing step.

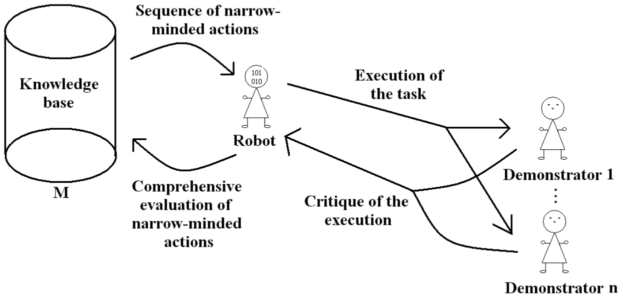
Knowledge base
M
Sequence of narrow-minded actions
101 010
Robot
Execution of the task
Demonstrator 1
⋮
Demonstrator n
Comprehensive evaluation of narrow-minded actions
Critique of the execution

Figure 2. The second stage of our procedure. In this step, the robot will refine the knowledge base M into a more universal perspective of the task. The critique of the demonstrators to the execution of the robot points out the pragmatic value of each behavior demonstrated in the 1^{st} stage of the procedure.

After each feedback is received by the robot from d_i, the confidence of each (p_m,a_n) used in the most recent execution is updated according to:

$$c := c + r_i * f(\,feedback\,),$$

where r_i is the demonstrator's credibility (as explained in the next subsection) and $f(\ feedback\ )$ is a function that depends also on the similarity search procedure used in the construction of the current task execution.[1]

It is noteworthy that the confidence c of each perception-action pair affects the similarity search procedure: it should result, for less trusted elements of M, in a smaller probability of being employed in future executions of the task.

Incentive-compatibility on the Critiquing Stage

It is clear that this desirable fusion of perspectives in M is strongly dependent on the quality of the critiquing signal. But unfortunately, the premises of our model make it natural to expect that demonstrators' spontaneous feedback would be correlated to their greedy executions demonstrated in step 1. In this case, the critiquing signal would not be very informative to a robot that already possesses the results of the original demonstrations.

This observation motivates the introduction of a reputation mechanism that attempts to control the human agents' behavior in the critiquing step. In this mechanism, we assign to each agent d_i a credibility rank r_i that estimates how much each agent's feedback to the robot's execution improves its performance on the task. The rationale for this design comes from the remark that each behavior profile for our task (e.g. driving defensively) is more appropriate in some situations and less desirable in others. Therefore, for each execution context we need a human input to indicate which profile would better fit the context. This is an incentive to prevent agents praising acts that resemble their own demonstrations and knocking other behaviors which might have come from fellow demonstrators. As a result, we induce this critique to be mindful. Therefore, the emerging pattern that results from this incentive-compatible scheme is the combination of the original demonstrated behaviors, normalized by their effectiveness in each context.

Since our assumption is that we don't know the utility functions of our agents, a natural question that comes up is how to evaluate the effect of a critique on the robot's execution in an acceptable way? To answer this question, we make use of the set of world aspects O. Since we assumed before that the agent's preferences over each aspect is comonotonic, we can now generate a Pareto ordering over the values of these world aspects that resulted from each task execution. For example, for any given simulated drive of the robot, we can compute the number of crashes (O_c) and the duration of the drive (O_d). If we compare the values of $(o_c,o_d) = \theta_1$ from the initial execution of the robot with $(o_c',o_d') = \theta_2$ from a subsequent execution after each the demonstrator's critique, we can define that the feedback resulted in an improvement if, and only if,

$$\forall O_i \in O \ \ \theta_2(O_i) \underline{\succsim} \theta_1(O_i) \quad and\ \exists O_j \in O \ \ \theta_2(O_j) \succsim \theta_1(O_j),$$

where $\theta(O_i)$ means the value of aspect O_i under θ. An analogous calculation yields a definition of feedbacks that result in decline.

Now that we introduced a fair procedure to judge critiques from humans, we can apply it to our reputation mechanism. This method assigns to demonstrators' reputation r_i an initial value of r^0, and after each critique from the demonstrator, we update his/her reputation estimate using the following rule:

$$r_i := r_i + \alpha * result(\,feedback\,),$$

where result(.) is a function that returns 1 if the feedback resulted in an improvement, -1 if it resulted in decline, and 0 otherwise. Here, α is a parameter of the model which indicates how fast the reputation of agents should increase or decrease on a single step.

Noticeably, this design of the reputation requires an attentive act by the agent on the critiquing phase. If a demonstrator adopts the strategy of persistently defending an ineffective behavior profile in detriment of giving truthful evaluations of the context presented by the robot, it is expected that the agent's reputation will drop to a point where it has no meaningful effect on the policy of the robot. Therefore, in order to continue influencing the result of the robot's policy (in other words, continue advocating for their own behavior profile), the agent must be mindful when evaluating current executions.

Related Work

As mentioned above, our model is an extension of (Argall, Browning and Veloso, 2007). In that work, the authors demonstrate how it may be possible to take advantage of contextual criticism by humans to teach a robot how to intercept a ball. Like our model, their framework also involves two stages. In the first, they assume that a single demonstrator presents the robot with a sequence of executions. Therefore, our model's first stage can be seen as a parallel instance of the original version, where in each new instance a particular demonstrator presents a set of executions to the robot.

Additionally, Argall et al. introduced the critiquing stage following the initial demonstration from the humans. In this phase, our departure is conceptually stronger. Even though both models assign to each pair (p_m,a_n) of a perceived and an action a measure of confidence, in our

[1] Argall et al. use a 1-NN as a similarity search procedure and the inverse of the distance between the actual perception and the actual execution point as f(feedback). The latter is to avoid penalizing decision pairs over contexts which they had weak correlation.

model this confidence parameter is more general. Not only does it represent a quantitative uncertainty, but we also endow its semantics as an intersection of diverging perspectives from different demonstrators. Finally, our reputation mechanism is not applicable to their non-strategic model.

In (Ekvall and Kragic, 2006), the authors introduce the possibility of having a group of humans cooperatively demonstrating a task to the robot. However, their model does not incorporate the non-cooperative behavior that might emerge when they explicitly cannot agree on goals or courses of actions.

Similar learning approaches of mixtures of cooperative sources can be found in the supervised learning literature. Product of experts (Hinton, 2000) and Mixture of experts (Jacobs et al., 1991) are examples of this trend.

Conclusions

We have introduced a model that suggests how a robot can learn from multiple demonstrators who disagree on the evaluation of the outcomes of a task. Our model makes weak assumptions over the preferences of the demonstrators, imposing only comonotonicity of preferences. We believe that a reputation mechanism is a sufficient element for inducing cooperative behavior from our demonstrators.

There are many ways in which we are expanding this research. First, we are working on the formalism of our problem and how to measure the quality of solutions. Since we do not assume any explicit representation on the utility of our demonstrators, we are looking for an objective measure to validate our claims of incentive-compatibility of the mechanism. Also, future experiments will help evaluate the quality of policies generated with our model, as well as allow comparison of our approach to other existing works. Ideally, we want our resulting policy to integrate the behavior of the different human agents in a coherent way. Another one open question is to measure the general satisfaction of the demonstrators with the solutions returned by our mechanism.

References

Argall, B., Browning, B., and Veloso, M. Learning by Demonstration with Critique from a Human Teacher. In *Proceedings of the Second Annual Conference on Human-Robot Interactions (HRI)*, Washington D.C., March 2007

Breazeal, C., and Scassellati, B. Robots that imitate humans. *Trends in Cognitive Sciences* Vol. 6 No. 11 November 2002

Ekvall, S., and Kragic, D. Learning Task Models from Multiple Human Demonstration. In *IEEE International Symposium on Robot and Human Interactive Communication*, 2006

Hinton, G. Training Products of Experts by Minimizing Contrastive Divergence. *Technical Report GCNU TR 2000-004, Gatsby Computational Neuroscience Unit, Univ. College London*, 2000

Jacobs, R., Jordan, M., Nowlan, S., and Hinton, G. Adaptive mixtures of local experts. *Neural Computation*, 3 79-87 1991