Qualitative Explanation of Controllers

Ivan Bratko, Dorian Šuc
University of Ljubljana, Faculty of Computer and Information Sc.
Ljubljana, Slovenia
{ivan.bratko, dorian.suc}@fri.uni-lj.si

Extended Abstract

We consider the following problem: Given a device, how does it work? How does it accomplish its purpose? In this paper we focus on controllers of dynamic systems. For example, given a controller for a dynamic system, like a crane or a plane, find an explanation of how the controller achieves its goal? We assume that the controlled dynamic system can be observed, either in reality or through simulation. We study two particular settings in which this problem arises:

(1) Reverse engineering of controller designs: there is a working system or a model that can be executed on a simulator, but the available documentation does not reveal the intuition behind the design. What are the basic ideas that led to the design of the artefact?

(2) Reconstruction of human operator's sub cognitive skill, also known as behavioural cloning: here control is done by a human operator who has the skill of controlling the system successfully, but cannot explain sufficiently well how he does it.

One approach to reverse engineer controllers of both types 1 and 2 above is by means of machine learning. A controller's execution traces are used as examples for machine learning, and a learning program aims at eliciting a useful description of the original controller. For the purpose of explaining how the controller works, it is essential that the learning system constructs meaningful symbolic descriptions that can be interpreted by the user. It is not sufficient to reproduce the control performance of the original controller, but to help the user's intuition to grasp the essential mechanism and causalities that enable the controller achieve the goal of control. Induction of so-called direct controllers in the form of regression trees has been used traditionally in behavioural cloning with some success, but also with clear drawbacks. Drawbacks are of various kinds: lack of robustness of induced controllers, and lack of explicit causalities, goals and subgoals that should feature in a good explanation of how the control strategy works.

In this talk, alternative approaches to the reconstruction of controllers from sample traces are discussed. The use of qualitative representations is advocated and their advantages compared to regression trees are analysed. We review our recent work along these lines. The concept of *indirect* controllers relies on inducing operator's *qualitative control trajectories* from his or her control traces. Qualitative trajectories can be obtained indirectly through qualitative abstraction of algebraic equations induced from numerical data, or directly by the QUIN learning program. QUIN induces *qualitative trees* from numerical data. Qualitative tree learning is similar to decision tree learning except that qualitative trees have qualitatively constrained functions in their leaves. QUIN has been applied to skill reconstruction and qualitative reverse engineering , and combined with QSIM-like qualitative simulation to generate dynamic explanation of an operator's control skill. Experiments with these

approaches in various domains will be presented, including behavioural cloning in the crane, acrobot and bicycle domains, and qualitative reverse engineering of an industrial crane controller and car suspension system. Two recent publications on this work are: I. Bratko, D.Šuc, Using machine learning to understand operator's skill, *Proc. IEA/AIE'02*, Cairns, Australia, 2002; D. Šuc, I. Bratko, Qualitative reverse engineering, *Proc. ICML’02 (Int. Conf. on Machine Learning)*, Sydney, Australia, 2002.