Engineering Applications of Artificial Intelligence 26 (2013) 652–659

Contents lists available at SciVerse ScienceDirect

Engineering Applications of Artificial Intelligence

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

An automated signalized junction controller that learns strategies by temporal difference reinforcement learning

Simon Box*, Ben Waterson

Transportation Research Group, Faculty of Engineering and the Environment, University of Southampton, SO17 1BJ, UK

ARTICLE INFO

Article history:
Received 6 October 2011
Received in revised form
19 February 2012
Accepted 20 February 2012
Available online 18 March 2012

Keywords:
Neural network
Reinforcement learning
Temporal difference
Traffic
Control
Junction

ABSTRACT

This paper shows how temporal difference learning can be used to build a signalized junction controller that will learn its own strategies through experience. Simulation tests detailed here show that the learned strategies can have high performance. This work builds upon previous work where a neural network based junction controller that can learn strategies from a human expert was developed (Box and Waterson, 2012). In the simulations presented, vehicles are assumed to be broadcasting their position over WiFi giving the junction controller rich information. The vehicle's position data are pre-processed to describe a simplified *state*. The state-space is classified into regions associated with junction control decisions using a neural network. This classification is the *strategy* and is parametrized by the weights of the neural network. The weights can be learned either through supervised learning with a human trainer or reinforcement learning by temporal difference (TD). Tests on a model of an isolated T junction show an average delay of 14.12 s and 14.36 s respectively for the human trained and TD trained networks. Tests on a model of a pair of closely spaced junctions show 17.44 s and 20.82 s respectively. Both methods of training produced strategies that were approximately equivalent in their equitable treatment of vehicles, defined here as the variance over the journey time distributions.

1. Introduction

1.1. Background

Urban signalized road junctions are usually controlled by active systems (e.g. Vincent and Peirce, 1988; Hunt et al., 1982), which use sensors to measure the state on the road. The state is then used by the control algorithm to inform decisions on which colour to set the traffic lights. Sensors such as inductive loops (Sreedevi, 2005) and microwave emitter/detectors (Wood et al., 2006) are commonplace and widely deployed in developed areas. New sensing technologies such as vehicle to infrastructure WiFi communications have been extensively investigated in recent years (Kompfner, 2008; COOPERS, 2010; SAFESPOT, 2010) leading to an Europe wide reservation of frequencies (IEEE 802.11p) for this type of communication.

The profusion of sensing technology leads to rich data that can be used for Urban Traffic Control (UTC). This enables the development of increasingly sophisticated control systems for signalized road junctions. In particular, data hungry *machine learning* algorithms can be employed to develop junction control systems that can learn improved strategies through various forms of training.

Recent important work on the optimisation of traffic signals has investigated a number of approaches including dynamic programming (Heydecker et al., 2007; Heung et al., 2005), genetic algorithms (Mikami and Kakazu, 1993), fuzzy-neural networks (Choy et al., 2003) and reinforcement learning (Chen and Heydecker, 2009). This work has shown how to use learning techniques to optimise parameters in signal control strategy or to select pre-defined strategies.

Here we are concerned with a pattern recognition approach where control decisions are made purely based on a classification of state space. Earlier work using this approach has shown how to use supervised learning to enable a junction controller to learn strategies from a human expert trainer (Box and Waterson, 2012). In this paper the approach is extended by the application of reinforcement learning to enable a junction controller to learn strategies through experience.

1.2. Context and motivation

Earlier work by the authors investigating the use of (vehicle transmitted) GPS + WiFi data in signal control has employed simulation to develop and evaluate control systems.

Under the *auctioning agent* control system (Waterson and Box, accepted for publication) the road network is discretized into regions and software agents monitoring each region calculate a

* Corresponding author.
E-mail address: s.box@soton.ac.uk (S. Box).

0952-1976/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2012.02.013

bid for priority. The bid is based on the positions and speeds of vehicles as reported over WiFi. At the junction a *junction agent* assigns the green light to those sections of road with the *highest bid*. Coordination between junctions is achieved through a *zone agent* which can re-weight bids to encourage coordination. In simulation tests the auctioning agent system using WiFi data outperformed the MOVA control system (Vincent and Peirce, 1988), which uses inductive loops.

The *human trained neural network* control system (Box and Waterson, 2012) uses the same system of bids as the auctioning agent system. However instead of using the bids as proxies for priority it uses the set of bids as an abstract simplified *state* describing the situation at the junction. A neural network is used to classify the resulting state space into decisions, namely which stage of the junction gets the green light. The training data for the network is provided by a human expert when they control the simulated junction via a computer game interface. Thus the human trained neural network is a machine learning junction control system that learns strategies from a human expert. In simulation tests the human trained neural network outperformed the auctioning agent control system.

The above control systems were both developed and tested on a simulation test-bed. This uses SIAS-Paramics micro-simulation software to simulate the movement of vehicles through the network. SIAS-Paramics is connected with a number of specially developed software modules to simulate sensor data, make control decisions and implement the control (i.e. change the traffic light colour) in the simulation. The same test-bed has been used in the research presented in this paper. It is described in full in Box and Waterson (2010a,b, 2012).

There are two principal shortcomings to using human experts to train machine learning junction controllers. Firstly, to implement this in practice would be costly because human time is relatively expensive. Secondly, the best possible performance of the system is limited to being as good as the human trainer. These shortcomings motivate the investigation into extending the *supervised learning* approach of the human trained neural network to build a *reinforcement* trained neural network.

As already described, junction control systems use measurements to determine the state on the road in order to make control decisions. However the state on the road right now tells us something about the decisions that were made in the past. In principal the controller can evaluate whether decisions made in the past were good or bad and learn from *these* data just as it learns from the data generated by the human trainer. This is the approach of *temporal difference* learning.

Research in other applications of artificial intelligence has shown that problems that can be solved using a neural network trained by supervised learning can also respond well to a neural network trained by temporal difference learning. A well known example of this is the work of Tesauro (2002) who developed the computer Backgammon program *Neurogammon*, which employed a neural network to learn strategies from human expert backgammon players. He then went on to develop "TD-gammon", a Backgammon program that used a neural network trained by temporal difference (TD) learning in simulations where the program competed against itself.

In this paper we present an adaptation to the neural network based junction control system described in Box and Waterson (2012). This adaptation enables the controller to be trained under simulation by temporal difference reinforcement learning. The principal contributions of the paper are as follows.

1. A new machine learning junction controller, which employs a two layer neural network to learn strategies through temporal difference reinforcement learning.
2. A comparison between the performance of the human trained junction controller and the TD trained junction controller in simulation tests.
3. Simulation tests of performance on both a simple isolated T-junction, and a pair of closely spaced junctions, where coordination is necessary.

2. Machine learning junction control strategies

2.1. Overview

2.1.1. Bids

When simulating GPS + WiFi data from vehicles we can collect estimates of the position and speed of every vehicle in the simulation. At any given time these data describe the *state* of the network. To make the problem more tractable and to speed up calculation time this raw description of the state is simplified in a pre-processing operation that generates *bids*.

To affect this the road network around the junction is divided into regions. Fig. 1 shows a schematic of one of the junctions discussed in this paper with the regions marked (a–d). Each region is monitored by an agent, which calculates a *bid* based on the position and speed data of the vehicles within that region. The bid is calculated using

$$B = \sum_{c \in C} 1 - \alpha V_c - \beta X_c \tag{1}$$

here C is the set of all vehicles monitored by the lane agent; V_c is the vehicle speed and X_c is the distance of the vehicle from the junction; α and β are the coefficients that can be tuned to adjust the relative influence that the number of vehicles, the vehicle speed and the vehicle distance each have on the size of the bid. In previous work (Box and Waterson, 2010b) it has been shown that (assuming S.I. units are used) values of $\alpha = 0.01\ \mathrm{sm}^{-1}$ and $\beta = 0.001\ \mathrm{m}^{-1}$ provide a good balance between influences. These values were adopted in this work.

The term "bid" is used because this method was first employed by the auctioning agents signal control algorithm (Waterson and Box, accepted for publication) where this bid was designed to be indicative of the need for priority on a section of road. For example more vehicles increase the bid, slower moving vehicles increase the bid and vehicles closer to the end of the road section increase the bid (and vice-versa). In the work presented here the set of bids from each of the regions is simply a description of the state of the

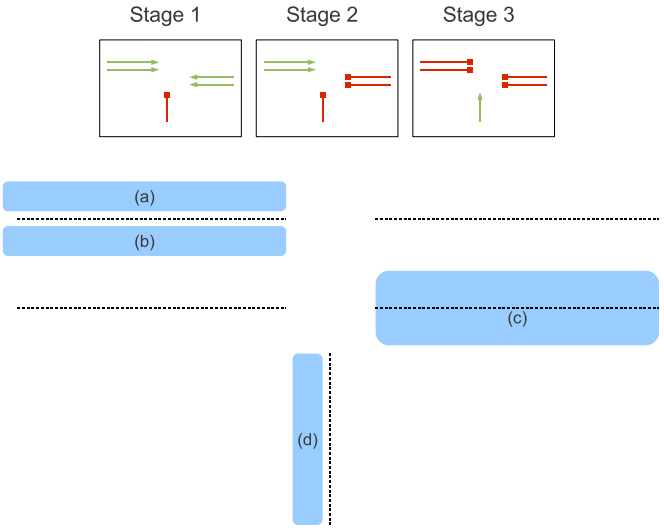

Fig. 1. Schematic of the Simple-T model showing the bid zones and the junction staging.