Transportation Research Part C 93 (2018) 179–197

Contents lists available at ScienceDirect

Transportation Research Part C

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

A reinforcement learning framework for the adaptive routing problem in stochastic time-dependent network

Chao Mao[a,c,*], Zuojun Shen[a,b,c]

[a] Department of Civil and Environmental Engineering, University of California Berkeley, Berkeley, CA 94720, USA
[b] Department of Industrial Engineering and Operations Research, University of California Berkeley, Berkeley, CA 94720, USA
[c] Tsinghua-Berkeley Shenzhen Institute (TBSI), Tsinghua University, Shenzhen 518055, China

ARTICLE INFO

Keywords:
Adaptive routing
Reinforcement learning
Q learning
Fitted Q iteration
Tree-based function approximation

ABSTRACT

Most previous work in addressing the adaptive routing problem in stochastic and time-dependent (STD) network has been focusing on developing parametric models to reflect the network dynamics and designing efficient algorithms to solve these models. However, strong assumptions need to be made in the models and some algorithms also suffer from the curse of dimensionality. In this paper, we examine the application of Reinforcement Learning as a non-parametric model-free method to solve the problem. Both the online Q learning method for discrete state space and the offline fitted Q iteration algorithm for continuous state space are discussed. With a small case study on a mid-sized network, we demonstrate the significant advantages of using Reinforcement Learning to solve for the optimal routing policy over traditional stochastic dynamic programming method. And the fitted Q iteration algorithm combined with tree-based function approximation is shown to outperform other methods especially during peak demand periods.

1. Introduction

In-vehicle route guidance system has become a more and more popular tool in people's daily lives, it can provide drivers with guidance on optimal routes from their current locations to predetermined destinations. Nowadays, many navigation devices can also receive real-time traffic information, which can be incorporated by the system to come up with smarter routing strategies. However, most of the current strategies are in a reactive fashion: when the link travel times are updated base on real-time traffic information, the system will recalculate the shortest path in the current network, and will recommend the new path to the user if it is faster. Since it does not take into account the possible realizations of link travel times in the future, this strategy is only suboptimal.

Different from a deterministic and static network where link travel times are fixed and do not change with time, the real traffic networks are usually stochastic and time-dependent due to the nature of periodicity and volatility of traffic demand. And it has been shown by Hall (1986) that the standard shortest path algorithms such as Dijkstra's algorithm and A* search fail to find the minimum expected travel time path in such networks. It was shown that the optimal "route choice" is not a simple path but an adaptive decision rule: an optimal successor node is chosen based on the arrival time of the current node, and further choices are made only when later nodes are reached. A method based on dynamic programming was proposed to find the optimal time-adaptive decision rule. It should be noted that Hall's adaptive routing model is a parametric model since the travel time probability distributions for all the links in the network are assumed to be known and utilized in the algorithm. Following Hall's work, a large number of studies have been conducted to address the adaptive routing problem in different settings, most of which are based on parametric models and are

* Corresponding author at: Department of Civil and Environmental Engineering, University of California Berkeley, Berkeley, CA 94720, USA.
E-mail address: chaomao@berkeley.edu (C. Mao).

https://doi.org/10.1016/j.trc.2018.06.001
Received 20 March 2017; Received in revised form 1 June 2018; Accepted 2 June 2018

summarized as follows.

1.1. Prior work: parametric models

Fu and Rilett (1998) extended the shortest path problem in dynamic and stochastic networks to the case where link travel times are defined as continuous-time stochastic processes. A probability-based formula for calculating the mean and variance of the travel time for a given path was developed and a heuristic algorithm based on the k-shortest path algorithm was proposed. Their model required the information on the mean and standard deviation of the link travel time as a function of the time of the day. Similarly, Miller-Hooks and Mahmassani (2000) presented two modified label-correcting algorithms for the problem of generating least expected time (LET) paths in stochastic and time-dependent networks. Travel times on the network were represented as random variables with probability distribution functions that vary with time. Fu (2001) also examined the adaptive routing problem in networks where link travel times were modeled as random variables with known mean and standard deviation, but the time-dependency of link travel times was handled with an algorithmic scheme. Based on the closed-loop routing policy in Fu (2001), Du et al. (2013) integrated traveller preferences in terms of travel time and travel time variability into the decision process. And they adopted a discrete distribution updated in real time to describe the dynamic characteristics of the link travel time.

While all the above works assumed that link travel costs are independent from each other, many other works have considered link-wise correlations in the stochastic networks. Waller and Ziliaskopoulos (2002) addressed the stochastic shortest path problem with recourse when limited forms of spatial and temporal arc cost dependencies were accounted for. One-step spatial dependence was assumed in such a way that if information from the predecessor arc was given, no further spatial information had an impact on the expected current arc cost. And this relationship was reflected in the conditional probability matrices. Also, Gao and Chabini (2006) studied routing policy problems in a general stochastic time-dependent network with both time-wise and link-wise dependency and perfect online information. A joint distribution of link travel times was used to represent the stochastic network, although it was difficult to be estimated in practice. Later, Gao and Huang (2012) expanded upon past research by examining the optimal routing problem with partial or no online information. A heuristic instead of an exact algorithm was designed and employed based on a set of necessary conditions for optimality. However, discrete distributions of link travel times were assumed for the convenience of defining routing policies. And the resulting algorithm was strongly polynomial in the number of support points, which might be exponential to the number of links in real networks.

A number of other researchers also attempted to model the stochastic temporal dependence of link costs using Markov chain. Psaraftis and Tsitsiklis (1993) examined the shortest path problem in acyclic networks in which arc costs are known functions of certain environment variables, and each of these variables evolves according to an independent Markov process. Azaron and Kianfar (2003) applied the stochastic dynamic programming to find the dynamic shortest path in stochastic dynamic networks, in which the arc lengths were independent random variables with exponential distributions. The parameter of the exponential distribution was assumed to be a function of the state of certain environmental variable, which would evolve in accordance with a continuous time Markov process. Later, Kim et al. (2005a) developed a decision-making procedure for determining the optimal driver attendance time, optimal departure times, and optimal routing policies under time-varying traffic flows based on a Markov decision process formulation. They assumed that each observed link can be in one of two states (congested or uncongested) that determined the travel time distribution used. However, when the number of observed links with real-time traffic information increases, the off-line calculations can be computationally intractable. To address this issue, Kim et al. (2005b) proposed a procedure for state space reduction. Taking into account the incident induced delays, Güner et al. (2012) proposed a stochastic dynamic programming formulation for dynamic routing of vehicles in non-stationary stochastic networks subject to both recurrent and non-recurrent congestion.

1.2. Motivation: non-parametric model-free methods

As can be seen in the above review, most of the previous work in addressing the adaptive routing problem has been focusing on developing parametric models to reflect the network stochasticity and designing efficient algorithms to solve these models. In their models, some finite set of parameters were used to represent the network characteristics, such as link travel time distributions, link correlations, or Markov processes. When applying them to real networks, we have to first estimate these parameters for the model based on some training data, and then solve for the routing policies based on the developed algorithms. There are many clear benefits of using parametric models: first, they are easy to understand and the results are more interpretable; second, the parameters can be learned quickly from a small set of data; and most importantly, they are more generally applicable, meaning that once the model is established for a certain network, we can solve for the best routing strategy from any origin to any destination quite efficiently. However, there are also some unavoidable limitations of the parametric models:

- First, in most of the parametric adaptive routing models we have seen, strong assumptions have to be made to allow for efficient solutions. However, these assumptions might not be consistent with the cases in real networks. For example, in Azaron and Kianfar (2003), link costs were assumed to be independent random variables with known exponential distributions, which might be difficult to validate in real networks since the distributions can vary a lot.
- Second, some of the proposed parametric models still suffer from the curse of dimensionality, i.e. they can be applied to small networks but cannot be incorporated into larger networks. For instance, the algorithm in Gao and Huang (2012) is polynomial in the number of support points of the discrete joint link travel costs distribution, which can be exponential to the number of links in the network. Thus in many cases, some approximation have to be applied to these models to allow for tractable solutions, which