Transportation Research Part C 72 (2016) 225–238

Contents lists available at ScienceDirect

Transportation Research Part C

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

A gradient boosting logit model to investigate driver's stop-or-run behavior at signalized intersections using high-resolution traffic data

Chuan Ding, Xinkai Wu *, Guizhen Yu, Yunpeng Wang

School of Transportation Science and Engineering, Beijing Key Laboratory for Cooperative Vehicle Infrastructure System and Safety Control, Beihang University, Beijing 100191, China

ARTICLE INFO

Article history:
Received 16 June 2016
Received in revised form 18 September 2016
Accepted 28 September 2016

Keywords:
Driving behavior
Gradient boosting logit
Variable importance
High-resolution traffic even data

ABSTRACT

Driver's stop-or-run behavior at signalized intersection has become a major concern for the intersection safety. While many studies were undertaken to model and predict drivers' stop-or-run (SoR) behaviors including Yellow-Light-Running (YLR) and Red-Light-Running (RLR) using traditional statistical regression models, a critical problem for these models is that the relative influences of predictor variables on driver's SoR behavior could not be evaluated. To address this challenge, this research proposes a new approach which applies a recently developed data mining approach called gradient boosting logit model to handle different types of predictor variables, fit complex nonlinear relationships among variables, and automatically disentangle interaction effects between influential factors using high-resolution traffic and signal event data collected from loop detectors. Particularly, this research will first identify a series of related influential factors including signal timing information, surrounding traffic information, and surrounding drivers' behaviors using thousands drivers' decision events including YLR, RLR, and first-to-stop (FSTP) extracted from high-resolution loop detector data from three intersections. Then the research applies the proposed data mining approach to search for the optimal prediction model for each intersection. Furthermore, a comparison was conducted to compare the proposed new method with the traditional statistical regression model. The results show that the gradient boosting logit model has superior performance in terms of prediction accuracy. In contrast to other machine learning methods which usually apply 'black-box' procedures, the gradient boosting logit model can identify and rank the relative importance of influential factors on driver's stop-or-run behavior prediction. This study brings great potential for future practical applications since loops have been widely implemented in many intersections and can collect data in real time. This research is expected to contribute to the improvement of intersection safety significantly.

1. Introduction

In recent years, concerns over intersection safety issue have prompted a growing body of study into the driver's Stop-or-Run (SoR) behavior at signalized intersections. There is a significant number of intersection crashes caused by the driving

* Corresponding author.
E-mail addresses: cding@buaa.edu.cn (C. Ding), xinkaiwu@buaa.edu.cn (X. Wu), yugz@buaa.edu.cn (G. Yu), ypwang@buaa.edu.cn (Y. Wang).

http://dx.doi.org/10.1016/j.trc.2016.09.016
0968-090X/

behavior every year. According to the National Highway Traffic Safety Administration's (NHTSA) report (NHTSA, 2012), more than 2.3 million crashes occurred at the intersection, and a great number of them were related to the drivers' SoR behavior such as yellow-light running (YLR) and red-light running (RLR). Therefore, it is critical to model and understand the drivers' SoR behavior at signalized intersections, and thereby predict it accurately to improve the intersection safety (Wu et al., 2013).

Many studies were undertaken to model and predict the YLR and RLR using the statistical regression models (e.g. Lu et al., 2015; Wu et al., 2013; Zhang et al., 2009; Bonneson and Son, 2003; Sheffi and Mahmassani, 1981; etc.). However, a critical problem for these models is that the relative influences of predictor variables on driver's SoR behavior were not evaluated. Understanding the relative importance of the influential factors on driver's SoR behavior would significantly help SoR prediction therefore contributes to the future improvement of intersection safety. Furthermore, understanding the relative importance of the influential factors would help to identify influential factors for which data should be collected and maintained. This is beneficial for data collection and maintenance as the process of collecting and maintaining data is cost-prohibitive. Although the sensitivity analysis can be conducted based on the statistical regression model (Saha et al., 2015), only one variable is evaluated at one time under the assumption that other variables remain the unchanged values. Therefore, the important relationships among the influential factors have been ignored.

To address the above mentioned challenge, this research proposes a brand new approach which applies a recently developed data mining approach called gradient boosting logit model (Zhang and Haghani, 2015; Saha et al., 2015). The proposed method is based on traffic data collected from loop detectors. Much existing research on YLR or RLR were using the high quality video data (e.g. Bonneson and Son, 2003; Bonneson and Zimmerman, 2004; Gates et al., 2007; Yang and Najm, 2007; Zhang et al., 2009; etc.). Video data is a reliable source, but such data is relatively rare since the data quality is constrained by the types of video cameras. Furthermore, most of the video data is off-line, and real-time video data analysis is time-consuming and costly.

The proposed research is based on loop detector data, which can be easily and automatically obtained in real time with low cost since most of signalized intersections have been equipped with loop detectors. Particularly, this research utilizes high-resolution traffic event data collected by loop detections. Traditional loop detector data which are usually aggregated into 30 s, 5 min or even 15 min are too coarse to describe individual driver's SoR behavior. With recent improvement of data collection methods (Lu et al., 2015; Liu et al., 2009; Smaglik et al., 2007), high-resolution traffic data (event-based or second-by-second data), which provide detailed vehicle arrivals and departures from loop detectors, become more and more popular. Such data, combined with the signal phase changes provided by signal control system, could be better used to analyze and predict the driver's SoR behavior (Wu et al., 2013).

The proposed data mining approach is designated to address challenging problems, such as drivers' behavior study, which has mixed types of predictor variables and complex nonlinear relationships. The proposed method can also automatically disentangle interaction effects between influential factors. To implement this approach, this research will first identify a series of related influential factors including signal timing information (e.g. time to yellow start and used yellow time), surrounding traffic information (e.g. occupancy time and time gaps of surrounding vehicles), and surrounding drivers' behaviors (i.e. drivers' RLR, YLR, and green-light running (GLR) decisions) using thousands drivers' decision events including YLR, RLR, and first-to-stop (FSTP) extracted from high-resolution loop detector data from three intersections. Then the research applies the proposed data mining approach to search for the optimal prediction model for each intersection. Furthermore, a comparison was conducted to compare the proposed new method with the traditional statistical regression model. In contrast to sensitivity analysis and other machine learning methods as 'black-box' procedures (Ding et al., 2015; Ma et al., 2015; Yu et al., 2014), the proposed method can identify and rank the relative importance of influential factors on driver's SoR behavior prediction. This study is expected to contribute to the improvement of intersection safety significantly.

The remaining of the paper is organized as follows. The second section provides a brief description of data collection, followed by the model specification. Model results and discussion are demonstrated in the forth section. Conclusion and future research directions are outlined at the end.

2. Data collection

Data collection has three parts: first is to collect high-resolution traffic and signal event data; second is to identify YLR, RLR and FSTP during yellow using stop-bar detectors, and third is to match events between stop-bar and advance detectors since the information collected from advance detectors will be used for investigation. Here stop-bar detectors are the detectors located right behind the stop-line and advance detectors are the detectors located 400 feet upstream from the stop-line (see Fig. 1).

2.1. High-resolution traffic and signal event data collection

The High-resolution traffic event data was collected by the SMART-SIGNAL (Systematic Monitoring of Arterial Road Traffic and SIGNAL) system developed at the University of Minnesota (Liu et al., 2009).The SMART-SIGNAL is capable of continuously collecting and archiving high-resolution event-based vehicle-detector actuation and signal phase change data. The