Transportation Research Part C 58 (2015) 308–324

Contents lists available at ScienceDirect

Transportation Research Part C

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

A gradient boosting method to improve travel time prediction

Yanru Zhang *, Ali Haghani

Department of Civil & Environmental Engineering, University of Maryland, College Park, MD 20740, United States

ARTICLE INFO

Article history:
Received 2 May 2014
Received in revised form 17 February 2015
Accepted 19 February 2015
Available online 29 March 2015

Keywords:
Short-term forecasting
Travel time
Ensemble learning
Machine learning
Gradient boosting regression tree
Random forest

ABSTRACT

Tree based ensemble methods have reached a celebrity status in prediction field. By combining simple regression trees with 'poor' performance, they usually produce high prediction accuracy. In contrast to other machine learning methods that have been treated as black-boxes, tree based ensemble methods provide interpretable results, while requiring little data preprocessing, are able to handle different types of predictor variables, and can fit complex nonlinear relationship. These properties make the tree based ensemble methods good candidates for solving travel time prediction problems. However, applications of tree-based ensemble algorithms in traffic prediction area are limited. In this paper, we employ a gradient boosting regression tree method (GBM) to analyze and model freeway travel time to improve the prediction accuracy and model interpretability. The gradient boosting tree method strategically combines additional trees by correcting mistakes made by its previous base models, therefore, potentially improves prediction accuracy. Different parameters' effect on model performance and correlations of input–output variables are discussed in details by using travel time data provided by INRIX along two freeway sections in Maryland. The proposed method is, then, compared with another popular ensemble method and a bench mark model. Study results indicate that the GBM model has its considerable advantages in freeway travel time prediction.

1. Introduction

Travel time, as an effective measure of freeway traffic conditions, can be easily understood by both travelers and transportation managers (Yeon et al., 2008). The provision of accurate travel time information through Advanced Traveler Information Systems (ATIS) enables travelers to make informed decisions for departure time, route, and mode choice, and therefore releases travelers' stress and anxiety. In addition, predicted travel time is valuable information for transportation managers in proactively developing Advanced Traffic Management System (ATMS) strategies. The Freeway Performance Measurement System (PeMS), the Split Cycle Offset Optimization Technique System, and the Sydney Coordinated Adaptive Traffic System are all successful traffic operation systems that either use travel time information as performance measures or as their module input (Choe et al., 2002; Yang et al., 2010). Apart from its important role in the successful implementation of intelligent transportation systems (ITS), travel time estimation and prediction are complex and challenging tasks. Resulting from the interactions among different vehicle-driver combinations, and exogenous factors such as weather, demand, and roadway conditions, travel time often experiences strong fluctuations across different periods and traffic conditions. These rapid fluctuations are often complex and difficult to predict. Fully understanding these fluctuations and developing accurate travel time prediction algorithms is critical.

* Corresponding author.

http://dx.doi.org/10.1016/j.trc.2015.02.019
0968-090X/

Inspired by the need of travel time predictions, a wide range of methodologies has been proposed in the literature, including historical average and smoothing (Farokhi Sadabadi et al., 2010; Smith and Demetsky, 1997; Williams et al., 1998), statistical and regression methods (Min and Wynter, 2011; Fei et al., 2011; Li et al., 2013b), various machine learning techniques (Zhang and Rice, 2003; Wang and Shi, 2012; Wei and Chen, 2012) and traffic flow theory based methods (Li et al., 2013a, 2014). For an extensive review of existing traffic prediction models refer to review papers (Vlahogianni et al., 2014; Van Lint and Van Hinsbergen, 2012). Among these methods, autoregressive integrated moving average (ARIMA) type model is widely recognized as an accepted framework to build freeway traffic prediction model, due to its well-defined theoretical foundations and effectiveness in prediction (Karlaftis and Vlahogianni, 2009). The ARIMA type model is gradually becoming a benchmark to compare with newly developed prediction models (Zhang et al., 2013b). There is a vast amount of freeway traffic prediction literature on such models (Kamarianakis and Prastacos, 2005; Min and Wynter, 2011; Zhang et al., 2013a). The ARIMA type or, more generally, statistical modeling approaches, assume certain model structures for the data. These models provide interpretable parameters with straightforward model structures. They can make very good predictions when traffic shows regular variations. Another type of promising prediction method, the machine learning (ML) algorithm, also generates great interest in traffic prediction filed. Some successful applications include: neural networks (Van Lint et al., 2005; Wei and Chen, 2012; Van Hinsbergen et al., 2009), support vector machines (Wang and Shi, 2012; Hong, 2011), and hybrid or ensemble techniques (Antoniou et al., 2013; Chen and Wu, 2012; Wang and Shi, 2012). In contrast to the statistical modeling approach, machine learning algorithm does not assume any specific model structure for the data but treat it as unknown. Therefore, it is more efficient in handling large data with any degree of complexity. The ML method can also capture the underlying structure of data even when they are not apparent (Vlahogianni et al., 2014). One of the disadvantages of most machine learning algorithms is the lack of interpretability, which limits their applications in traffic prediction.

In recent years, ensemble based algorithms reached a celebrity status in solving prediction and classification problems. They have been applied to different fields and have achieved great success (Zhou, 2012). The $1 Million Netflix Prize competition is a famous example: the winning team ensembles different algorithms to predict user rating for films, which produces the best accuracy among all participates (Koren, 2009). Within all different ensemble methods, the tree-based ensemble method is a popular one. Instead of fitting a single "best" model, the tree-based ensemble method strategically combines multiple simple tree models to optimize predictive performance. Drawing on insights and techniques from both statistical and machine learning methods (Elith et al., 2008), they not only achieve strong predictive performance, but also identify and interpret relevant variables and interactions. Interpretability of the tree-based ensemble model enables transportation decision makers to better understand the output of the model and is critical in analyzing relationship between traffic and their influential factors. In addition, the tree-based ensemble method can handle different types of predictor variables, requires little data preprocessing, and can fit complex nonlinear relationship (Elith et al., 2008). These properties make the tree-based ensemble methods good candidates in solving transportation problems, such as traffic prediction and incident classification.

However, there are limited studies on the application of tree-based ensemble methods in transportation field. Leshem and Ritov (2007) combined the random forests algorithm into Adaboost algorithm as a weak base model to predict traffic flow. The proposed algorithm is proved to be able to deal with missing data and is effective in predicting multiclass classification problems. Hamner (2010) applied random forest in travel time prediction and is one of the winners for the IEEE ICDM Contest: TomTom Traffic Prediction for Intelligent GPS Navigation. Their proposed method outperforms other models in terms of prediction accuracy. Wang (2011) applied an ensemble bagging decision tree (ensemble BDT) to predict weather impact on airport capacity and demonstrated the superior performance of ensemble BDT compared with single support vector machine classifier. Ahmed and Abdel-Aty (2013) utilized a stochastic gradient boosting method in identifying hazardous conditions based on traffic data collected from different sensors. Their study results suggested that the proposed stochastic gradient boosting method has considerable advantages over classical statistical approaches. Similarly, Chung (2013) applied boosted regression trees to study crash occurrence. Both studies utilized the boosting method to classification problems. To the best of our knowledge, research on gradient boosting tree in travel time prediction has not been fully documented to date.

This study proposes a tree-based ensemble method to predict travel time on a freeway stretch by considering all relevant variables derived from historical travel time data. Belonging to the machine learning category, the tree-based ensemble methods often have superior prediction performance over classical statistical ones. Driven by the successful application of random forest in traffic parameter prediction, a gradient boosting tree-based travel time prediction method is proposed to uncover hidden patterns in travel time data to enhance the accuracy and interpretability of the model. Different from the random forest algorithm that averages a large collection of trees from random sampling (Breiman, 2001), the gradient boosting method sequentially generates base models from a weighted version of the training data to strategically find the optimal combination of trees. Each step of adding another base model is aimed at correcting the mistakes made by its previous base models. Therefore, the gradient boosting method has the potential to provide more accurate predictions.

The rest of this paper begins with a detailed description of tree-based ensemble methods from the statistical perspective, including explanations of the single regression tree, the random forest and the gradient boosting tree or the gradient boosting method (GBM). The next section discusses the application of the GBM method in freeway short-term travel time prediction, which includes details of data preparation, model optimization, interpretation, and comparison. Discussion and conclusion are outlined at the end.