Transportation Research Part C 58 (2015) 292–307

Contents lists available at ScienceDirect

Transportation Research Part C

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

Robust causal dependence mining in big data network and its application to traffic flow predictions

Li Li [a,b], Xiaonan Su [a], Yanwei Wang [a,c], Yuetong Lin [d], Zhiheng Li [a,c,*], Yuebiao Li [a]

[a] Department of Automation, Tsinghua University, Beijing 100084, China
[b] Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, SiPaiLou #2, Nanjing, China
[c] Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
[d] Department of Electronics & Computer Engineering Technology, Indiana State University, IN 47809-9989, USA

ARTICLE INFO

Article history:
Received 1 May 2014
Received in revised form 16 February 2015
Accepted 3 March 2015
Available online 8 April 2015

Keywords:
Big Data
Traffic flow prediction
Causal dependence
Lasso regression
Robust

ABSTRACT

In this paper, we focus on a special problem in transportation studies that concerns the so called "Big Data" challenge, which is: *how to build concise yet accurate traffic flow prediction models based on the massive data collected by different sensors*? The size of the data, the hidden causal dependence and the complexity of traffic time series are some of the obstacles that affect making reliable forecast at a reasonable cost, both time-wise and computation-wise. To better prepare the data for traffic modeling, we introduce a multiple-step strategy to process the raw "Big Data" into compact time series that are better suited for regression and causality analysis. First, we use the Granger causality to define and determine the potential dependence among data, and produce a much condensed set of times series who are also highly dependent. Next, we deploy a decomposition algorithm to separate daily-similar trend and nonstationary bursts components from the traffic flow time series yielded by the Granger test. The decomposition results are then treated by two rounds of Lasso regression: the standard Lasso method is first used to quickly filter out most of the irrelevant data, followed by a robust Lasso method to further remove the disturbance caused by bursts components and recover the strongest dependence among the remaining data. Test results show that the proposed method significantly reduces the costs of building prediction models. Moreover, the obtained causal dependence graph reveals the relationship between the structure of road networks and the correlations among traffic time series. All these findings are useful for building better traffic flow prediction models.

1. Introduction

The operation and management of contemporary Intelligent Transportation Systems (ITS) relies on a myriad of sensors/actuators to generate important data and actions. The magnitude of data has increased dramatically over the years as the result of growing number of sensing and actuating units in the ITS, and it poses a serious challenge on how to effectively utilize these "Big Data" to aid traffic research and practices. Unfortunately, traditional databases and search engines can only offer functions such as data storage, indexing and query, and are incapable of turning the "Big Data" into applicable knowledge or building inference mechanism to quantify the degree of support that data offers for decision making.

[*] Corresponding author at: #822, Central Main Building, Tsinghua University, Beijing 100084, China. Tel.: +86 (10) 62795503.
E-mail address: zhhli@tsinghua.edu.cn (Z. Li).

http://dx.doi.org/10.1016/j.trc.2015.03.003
0968-090X/© 2015 Elsevier Ltd. All rights reserved.

One of the critical tasks in ITS is traffic predictions, which has been heavily investigated by multiple researchers (Smith et al., 2002; Vlahogianni et al., 2004; Chrobok et al., 2004). Since it is generally believed that more data help to build better prediction models (Bhaskar et al., 2011; Van Lint and Hoogendoorn, 2010; Heilmann et al., 2011), there have been numerous attempts that target mining the available data generated by all traffic sensors and actuators (Zhang et al., 2011; Faouzi et al., 2011).

However, with the emergence of "Big Data", in particular considering some ITS systems are already generating gigabytes, and on the verge of generating terabytes of data per day or even beyond, it has become apparent that feeding all data into one prediction model becomes both computationally prohibitive and counter-productive.

A potential solution is to employ parallel computing techniques where the computational tasks are divided into a few smaller subtasks to be run on different computers separately (Chen et al., 2013). However, the benefit-to-cost ratio of this kind of approach remains to be further examined. While it seems intuitively obvious that more data can aid prediction, the performance of prediction often may not be dramatically improved in practice.

Another more practical method is to find and use the most relevant data to build parsimonious prediction models. To this end, different methods have been proposed in the last decade, including the hierarchical fuzzy model (Stathopoulos et al., 2010) that fuses multivariate data for traffic flow prediction, the adaptive Lasso method (Least Absolute Shrinkage and Selection Operator) that finds the most influential data in building prediction models (Kamarianakis et al., 2012), and the Graphical Lasso method (Sun et al., 2012) that builds sparse Bayesian Network models for prediction. As summarized in Vlahogianni et al. (2014), the main challenge of these studies lies in how to correctly retrieve both temporal characteristics and spatial dependence of traffic flow time series collected at different locations. The size of the data, the hidden causal dependence and the complexity of traffic time series all hinder us to reach this goal.

In this paper, we focus on this special problem, which is: *how to build traffic prediction models based on the massive data collected by different sensors*?

Particularly, the first problem that we need to address is: when we are discussing dependence, how to thoroughly classify and well define the so called "*causality*" between different sources of data?

In order to answer this question, we turn to the famous Granger causality test. The search for causal relationship dates back to a millennium ago due to its indubitable importance and wide applications. Since there lacks a uniform yet intuitive understanding of cause and effect, the causal relationship is often difficult to define for complex systems. In 1956, Norbert Wiener proposed that one variable (usually appears as a particular time series) could be called 'causal' to another variable if the ability to predict the second variable can be noticeably improved by incorporating information about the first variable (Wiener, 1956). In 1969, Clive Granger designed a practical implementation of this idea by checking the information gain of different linear autoregressive models for stochastic processes (Granger, 1969, 1980). Since then, Granger causality test has become a well-established method for identifying potential causal connectivity (Otter, 1991; Hlaváčková-Schindler et al., 2007; Bressler and Seth, 2011). It has gained tremendous success across many domains (Asimakopoulos et al., 2000; Lozano et al., 2009).

In this paper, we adopt the newly developed Lasso method based Granger causality model (Arnold et al., 2007; Lozano et al., 2009) to retrieve the most important causal relationships from the massive traffic data. However, the original Lasso method is based on the assumption of stationary time series and may fail when this assumption is not met. Therefore, the second problem we need to address is: how to simultaneously cope with the causal dependence modeling and the non-stationarity of traffic time series?

Realizing that traffic flow time series usually contain various components and are not strictly stationary (Chen et al., 2012; Zeng and Zhang, 2013; Zhang et al., 2014), we apply the decomposing technique that we had developed in the last decade (Li et al., 2014b) to classify the traffic data from each sensor into three kinds of patterns: the intra-day trend, the Gaussian type fluctuations, and the bursts. From the viewpoint of traffic prediction, these patterns represent two opposing aspects of the traffic series: the intra-day trend reflects the endogenous, self-driven and time-invariant characteristics; while the residual series, which includes fluctuations and bursts, reflects the exogenous, environment-dependent and time-variant characteristics. As already shown in Chen et al. (2012), building prediction models on the residual time series may significantly improve prediction accuracy. And in this paper, we further demonstrate that residual series also play a fundamental role in uncovering the underlying causal relationships in traffic series.

To this aim, we may use the Lasso algorithm to quickly filter out most of the unrelated data from the residual time series. However, since the bursts usually represent the impact of un-modeled environmental disturbance that may bias the estimation of standard Lasso algorithm, a robust Lasso algorithm is designed instead to remove the bursts first and then recover the dependence among the remaining time series. Test results show that the proposed method is able to appropriately recover both the temporal characteristics and spatial dependence of the original traffic flow time series. Moreover, the obtained causal dependence graph reveals the relationships between the structure of road networks and the correlations among traffic time series. All these findings help to build better prediction models.

Fig. 1 shows the five steps (models) to handle the interlaced difficulties. The rest of this paper will present each solution model respectively. Section 2 first explains how to decompose traffic flow time series for further study. Section 3 briefly reviews the theory of Granger causality models and then explains how to determine the causal dependence for traffic prediction. Section 4 presents the test results of this new method and discusses the relations between causality modeling and prediction modeling. Finally, Section 5 gives the conclusions.