Ding et al. Journal of Cloud Computing: Advances, Systems and Applications (2020) 9:25
https://doi.org/10.1186/s13677-020-00170-1

Journal of Cloud Computing: Advances, Systems and Applications

RESEARCH Open Access

An ensemble-learning method for potential traffic hotspots detection on heterogeneous spatio-temporal data in highway domain

Weilong Ding[1,2*], Yanqing Xia[1,2], Zhe Wang[1,2], Zhenyu Chen[3,4*] and Xingyu Gao[5*]

Abstract

Inter-city highway plays an important role in modern urban life and generates sensory data with spatio-temporal characteristics. Its current situation and future trends are valuable for vehicles guidance and transportation security management. As a domain routine analysis, daily detection of traffic hotspots faces challenges in efficiency and precision, because huge data deteriorates processing latency and many correlative factors cannot be fully considered. In this paper, an ensemble-learning based method for potential traffic hotspots detection is proposed. Considering time, space, meteorology, and calendar conditions, daily traffic volume is modeled on heterogeneous data, and trends predictive error can be reduced through gradient boosting regression technology. Using real-world data from one Chinese provincial highway, extensive experiments and case studies show our methods with second-level executive latency with a distinct improvement in predictive precision.

Keywords: Spatio-temporal data, Traffic trends, Ensemble learning, Highway, Big Data

Introduction

With the boom of inter-city transportation, highway plays an important role in modern urban life, and traffic congestion issue has become one of the most serious problems worldwide. In fact, the capacity of road network has not been explored enough [1], and vehicle guidance is imperative for officials [2] to reduce transportation risk. As a typical routine analysis in domain against that problem, *traffic trend* is to predict traffic volumes at toll stations in next few days, and significant for business to alleviate traffic congestion by dispersing traffic flow accordingly in highway network. It has been widely adopted to find *potential hotspots* of smart cities [3]. For such analyses, heterogeneous data from multiple sources can be employed integrally in domain nowadays. For example, toll data from toll stations keeps time-stamps and locations when a vehicle was entering or exiting a station, and has the advantages of exact locality and higher quality [4]. Besides, meteorological data and calendric data from dedicated or public service are also required. Implicit traffic patterns would appear periodically under specific weather conditions (e.g., heavy snow or rain) or on different days (e.g., traditional holidays). Such types of data are typical spatio-temporal, because both time and space attributes have to be considered integrally.

Such data can reflect not only current situation but also future trends [5] accordingly. However, it faces challenges to detect potential traffic hotspots through trends prediction due to inherent limitations in practice. On one hand, it is hard to hold executive performance like

* Correspondence: dingweilong@ncut.edu.cn; czy9907@gmail.com; gxy9910@gmail.com
[1]School of Information Science and Technology, North China University of Technology, Beijing 100144, China
[3]Big Data Center, State Grid Corporation of China, Beijing 100031, China
[5]Institute of Microelectronics, Chinese Academy of Sciences, Beijing 100029, China
Full list of author information is available at the end of the article

low latency when data grows into a huge size. Classic statistical models only work well on limited samples at given spatial locations, which are not suitable for trends prediction on massive data from multi-sources. For example, ARIMA regression treats sensory data as time-series, and predicts trends once at a single location [6]. Its predictive effect distinctly falls when data size grows. On the other hand, traditional methods cannot gain sufficient precision because they cannot fully cover key characteristics synthetically. For example, time restricted models applied in many scenes only emphasize temporal feature using insufficient business factors. In highway domain, besides temporal patterns from time-series, calendric impacts, spatial feature of locality (e.g., road network topology) and weather conditions in specific geographic ranges would also influence traffic trends directly. All of them have been seldom considered simultaneously in current works yet. In brief, it is not easy to find potential traffic hotspots through trends prediction in highway domain, in views of both effectiveness and efficiency.

In this paper, an ensemble-learning based prediction method is proposed for potential traffic hotspots detection in highway domain. Our contributions can be concluded as follows. (1) To adequately describe features on heterogeneous data, daily traffic volume is modelled fully considering time, space, meteorological condition, and calendric fact. Such general characteristics are necessary for trends prediction. (2) To improve performance and precision of daily trends prediction for traffic hotspots detection, an ensemble-learning model through gradient boosting regression is built. With second-level executive latency, it can reduce predictive error about nearly 10 percentages than traditional ways. (3) Having been evaluated in a practical scene, our work has convincing benefits on the real-world data in extensive experiments and case studies.

The rest of this paper is organized as follows. Section 2 shows background including motivation and related works; Section 3 elaborates our prediction method for traffic hotspots detection; Section 4 demonstrates performance and effects from experiments and case studies; Section 5 summarizes conclusion.

Background

Motivation

Our work originates from *Highway Big Data Analysis System* in Henan which is the most populated province in China. The system we built has been in production since October 2017 and is expected to improve routine business analytics for highway management through Big Data technologies. Operated by officers of *Henan Transport Department*, a billion records of heterogeneous data in recent 2 years have been imported into the system. Take toll data for example. A record of toll data has the structure in Table 1, which is typical spatio-temporal and contains 12 attributes including six entity attributes, two temporal attributes and four spatial attributes. Besides, other types of data, such as daily meteorological condition, solar and lunar calendar, and real-time license plate recognition data, are loaded into that system.

As one significant business analytics in highway domain, *potential traffic hotspots detection* is to evaluate and sort future traffic volumes. Here, *traffic volume* counts vehicles passing a spatial point during given time slots, and its prediction reflects *traffic trends* in different granularities. The *potential traffic hotspots* focused in this paper are the toll stations with top-k predictive daily traffic volumes. With traditional ways to predict traffic trends for the detection, toll data from sensors would be loaded into a data warehouse at the end of a day; after ETL (Extract, Transform, Load) step with necessary preprocessing [7], business OLAP (Online Analytical Processing) would be triggered to execute in the data warehouse; when completed, predicted results are available for business statistics and risk management. However in practice, such prediction brings long delays (e.g., 1 week) to release official reports due to complex processing procedure. Moreover, traditional models widely used in domain like classic ARIMA and wavelets do not fit well on huge data from hundreds of stations, because their predictive errors are only qualified on limited samples at a single location [8]. Accordingly, a novel solution is required for potential traffic hotspots detection on massive spatio-temporal data to improve both latency and precision. It is just our original motivation for practical transportation management.

Related work

For data analytics in specific domain, many works aim to handle massive data with Big Data [9]. Potential

Table 1 Structure of toll data

Attribute	Notation	Type
collector_id	toll collector identity	Entity
vehicle_license	vehicle identity	Entity
vehicle_type	vehicle type	Entity
card_id	vehicle passing card identity	Entity
etc_id	vehicle ETC card identity	Entity
etc_cpu_id	ETC card chip identity	Entity
entry_time	vehicle entry timestamp	Time
exit_time	vehicle exit timestamp	Time
entry_station	identity of entry station	Space
entry_lane	lane number of entry station	Space
exit_station	identity of exit station	Space
exit_lane	lane number of exit station	Space

hotspots detection is significant nowadays in many business domains, but still faces challenges in efficiency and precision. We would divide related works into two technical perspectives from hotspots detection and traffic trends prediction.

From the first perspective of hotspots detection, many solutions are proposed in domains like environmental monitoring, mobile communication, and transportation security analytics. It is imperative to find hotspots in modern urban life for better risk management and emergency response. On traffic signal data and air quality monitoring data in southern Brazil, relation analysis model is proposed [10] to identify pollution hotspots with main factors. By mapping spatial distribution of black carbon and PM2.5 to focused regions, strong relationship is observed with junctions managed by traffic signals. However, as a traditional statistical way, such model cannot hold effectiveness when huge data is imported. For transportation risk management, a Tobit model is developed [11] to identify potential hotspots of crash incidents. By relating urban grid-cell and influence factors, crash costs of pedestrian, including injury severity and exposure indicators, are modelled and ranked through Big Data technology. Although safety improvement is achieved, the solution as a post analytic on incidents data is not suitable for accumulated data condition (e.g., toll data in highway). Continuous response is often required for routine and crucial business like situation risky management. In analogous idea, an expert system is built [12] to visualize potential areas with a high density of incidents or risky situations. Besides inherent limitations discussed above, centralized data storage for its global repository restricts analytical performance due to non-scalable physical capacities. To find potential congestion hotspots, a complex network model is presented [13] to forecast scenarios and improve traffic control. But as an idealized concept model, it is hard to implement with interpretability and cannot be adapted instantly under fluctuant traffic conditions. In fact, hotspots detection through predictive volume trends is popular in mobile communication domain. Like [14, 15], traffic hotspots of mobile network are found among urban cells by dynamic data volume distribution. Although their queuing models are not appropriate in highway domain due to different business semantics, the basic idea of evaluating volume trends inspires our work in this paper. By predicted traffic trends on heterogeneous spatio-temporal data, a detection method for potential traffic hotspots in highway domain is proposed in this paper, which can hold low latency and low predictive errors than traditional ways.

From the second perspective of traffic trends prediction, current works can be classified into three types, each of which has its own advantages and disadvantages. First, by exploiting combinatorial dependencies among multivariable factors, linear models could improve predictive precision [1]. For example, through time series model, temporal characteristics of traffic volume are studied for traffic trends prediction [16]. By referring a time lag function, an extended ARIMA model is defined [17] to improve predictive accuracy on the traffic volume data. However, such offline methods would be disturbed by imperfect records especially for continuous data. In a real-time scene, extended Kalman-filtering model [18] and ARIMA+ [6] are employed for macroscopic and single-location traffic trends prediction respectively. Due to computational complexity with vectors or matrixes, such methods require too much time to tune parametric weights to guarantee executive performance. Second, non-linear models are also widely investigated due to its flexible feature presentation. To improve accuracy, wavelet model [19] and time delay neural network through genetic algorithm [20, 21] are used for traffic volume prediction. But they focus on short-term trends within 30 min and cannot be applied directly for longer periods like 1 day due to different temporal granularity. Moreover, most of them could only achieve values for limited locations in a prediction execution. Third, knowledge discovery methods through machine learning are popular owing to their high precision. Based on decision tree, a model is proposed [22] to estimate travel-time in Prague on surveillance camera data. It is suitable for data condition in low density, but its calculation with multiple scans and sorts would require too long latency to endure. Modern deep learning algorithms, such as Spatio-Temporal Semi-Supervised Learning (ST-SSL) [23], Residual Recurrent Graph Neural Networks [24], Convolutional Neural Network [25], and Attention-based Graph Convolutional Network [26], are also presented with pretty good predictive precision. They achieve good results in some fields like multi-media [27], however two main limitations restrict their wide adoption in practice for highway domain. On the one hand, such black-box models lack business interpretation for business technicians [4]. On the other hand, their executive latency, especially during training phase, is too long for instant response in highway management. As a machine learning model, our work holds high predictive precision. Besides, it promotes the model's interpretability to build a feature vector of daily traffic trends, considering multiple business characteristics from heterogeneous data. Meanwhile, latency of a prediction execution is second-level and feasible enough in practical highway.

In brief, potential traffic hotspots detection in highway domain still faces challenges to guarantee low latency and high predictive precision on massive data. Accordingly, we introduce our novel detection method fully

considering spatio-temporal characteristics on heterogeneous data with efficient trends prediction algorithm.

Traffic hotspot detection through daily trends prediction

Methodology

In highway domain, potential traffic hotspots can be found by evaluating recent traffic trends. Traffic trend could be depicted in different temporal or spatial granularity, such as that of short-term, long-term, single section, multiple sections, and so on. In this paper, we focus on traffic hotspots through daily trends defined below.

Definition 1:Traffic trends of daily traffic volume. The traffic trends are presented as the daily predictive traffic volumes at specific locations L in next few days D. Here, L is the set of toll stations in highway network, and D is a set of consecutive dates in future since the current day d_0.

Accordingly, potential traffic hotspots can be defined as follows.

Definition 2: Potential traffic hotspots. Based on Definition 1, the potential traffic hotspots in a day $d \in D$ is the locations L' which have top-K traffic volumes in descending order from traffic trends at any location $l \in L$ in d. Here, $K \in \mathbb{Z}^+$, $L' \subseteq L$.

Our method for potential traffic hotspots through trends prediction is proposed as Fig. 1. Four main parts are included in this framework.

Input data layer loads required data in for hotspots detection. Raw online *toll records* are received continuously through a message broker, and then aggregated as *traffic volume data* into No-SQL database. The preprocessing including data cleaning and aggregative calculation can be referred in our previous works [4, 28]. External calendar and meteorology data are extracted periodically from Web API of dedicated data sources, and then stored into a relational database. Besides, business basic data, such as profiles of station, section and highway line, has been imported in that relational database. On such heterogeneous data, the feature of traffic volume is modeled by **feature modeling** module. With the feature, **hotspot detection** module adopts GBRT (Gradient Boost Regression Tree) technology to build an ensemble-learning model after algorithmic parameters tuning. Daily traffic trends would be predicted by the trained model, written into the relational database, and used to find potential hotspots after sorting procedures. The visualized potential hotspots would be presented in an online map of **application** module.

In practice, our method as a routine analysis would be triggered to execute once a day at 12:00 a.m. on the data of recent 4 months, and output traffic trends for next 30 days. In each of those days, potential hotspots can be found interactively. In fact, although discussed in a specific domain, our method would be general in many scenes where recent trends of aggregative values (e.g., traffic volume here) are to be predicted on heterogeneous spatio-temporal data (e.g., toll data, meteorological data, and calendric data here).

Feature modelling

As the framework of our method in Fig. 1, potential hotspots are found by evaluating the trends of traffic volumes. Accordingly, traffic volume is the key ingredient for estimation and has to be modelled properly. Because traffic is always bi-direction at a toll station in highway

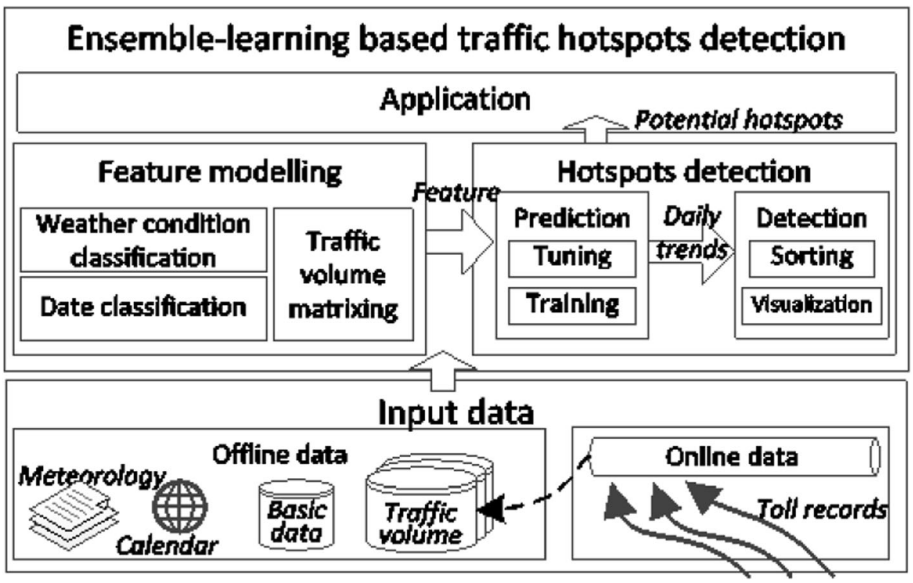

Fig. 1 Framework of our method

network, daily traffic volume can be formally defined as follows, inspired by the spatio-temporal attributes in Table 1.

Definition 3: Daily exit (entry) traffic volume. Daily exit (entry) traffic volume presented as xTF_l^d (nTF_l^d) counts the volume of vehicles V exiting (entering) toll station $l \in L$ in a day d.

Traffic volume is the exit one by default if no other emphasis. The exit traffic volume rather than the entry one is the main focus in domain because tolls would be charged only when vehicles exiting a station. From elaborative observation, we found the traffic volumes of a certain station correlate with dimensions of time (e.g., recent values at this station), space (e.g., values at adjacent stations), date (e.g., workday or not), and weather (e.g., extreme condition or not). Accordingly, the feature vector of exit traffic volume can be defined then.

Definition 4: Feature vector of exit traffic volumes. In any day $d \in D$, the feature of exit traffic volumes is a vector $V^d = (X^d, Y^d)$. $Y^d = (xTF_1^d, xTF_2^d, \ldots, xTF_l^d, \ldots, xTF_L^d)$ is the exit traffic volumes at any toll station $l \in L$. $X^d = (X_1^d, X_2^d, \ldots, X_l^d, \ldots, X_L^d)$, whose component is the characteristics at l in continuous days: $X_l^d = (W_l^d, D_d, xTF_l^{d-1}, \ldots, xTF_l^{d-i}, \ldots, xTF_l^{d-\theta+1}, nTF_{l'_1}^{d-1}, \ldots, nTF_{l'_j}^{d-1}, \ldots, nTF_{l'_{\eta-1}}^{d-1})$, $i = 0..\theta\text{-}1$, $j = 1..\eta\text{-}1$, $\theta \in \mathbb{Z}^+$, $\eta \in \mathbb{Z}^+$. Four dimensions exist in X_l^d: the first $W_l^d \in \{0, 1\}$ is weather category at location l in day d; the second part $D_d \in \{0, 1, 2\}$ is date category of day d; the third is exit traffic volumes at l in previous θ days; the fourth is entry traffic volumes in day d-1 at η-1 upstream dependent stations l'_1, l'_2, ..., $l'_{\eta-1}$. These factors are illustrated in details below.

(1). In the first meteorological dimension, extreme weather condition must be distinguished from the normal ones on meteorology data. In extreme weather condition, traffic in highway would be strictly controlled even closed by officials considering drivers' safety. Raw meteorology data of counties and cities has been loaded as the form W_l^d (*visibility*, *rain*, *temperature*(*low*, *high*), *wind*, *snow*) at $l \in L$ in d. According to domain standard [29], extreme weather (i.e., W_l^d =1) can be defined as follows.

$$W_l^d = \begin{cases} 1, \text{if}((visibility|rain|snow).level \geq 2) \\ \quad \text{or}((wind|temperature).level \geq 3), \\ 0, \text{otherwise}. \end{cases}$$

(2). In the second calendric dimension, holidays and weekends are differentiated from others according to solar and lunar calendars. In such days, traffic at stations of tourist areas would burst because much more private cars pass for tourism. Accordingly, for a given date d, D_d is presented as follows.

$$D_d = \begin{cases} 1, \text{if } d \text{ is a weekend}, \\ 2, \text{if } d \text{ is a holiday}, \\ 0, \text{otherwise}. \end{cases}$$

(3). In the third temporal dimension of X^d, traffic volumes in recent θ-1 days at station l are considered, which reflects time series characteristic. In the temporal dimension of Y^d, traffic volumes of day d at any location in are concerned.

(4). In the spatial dimension of X^d, entry traffic volumes in day d at η-1 upstream dependent stations are considered. Here, an ***upstream dependent location*** of l is such a station l'_j whose entry traffic volume influences the exit one at l. It can be found in ascending order by sorting absolute differences between cartographic distance (l, l'_j) and vehicles' average mileage at l. It just depicts the spatio-temporal correlation in trips from l'_j to l according to highway network topology. The details can be found in our previous work [4].

For consistent presentation, the ranges of all the scalar values have to be normalized as [0..1] by common Mean Normalization. Such normalization can reduce the interference by the dominating features with large range.

Trends prediction and potential hotspots detection

To predict traffic trends through the features we defined, an ensemble-learning model based on GBRT (Gradient Boosting Regression Tree) is designed in this section. GBRT combines weak models into a stronger one by iterative stages to improve predictive accuracy with an arbitrary loss function.

According to business habits for evaluation, least-square regression is adopted as loss function *Lf*, where the goal is to find a model F by minimizing mean squared error $\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)$. Here, following Definition 4, for an input x (i.e., a feature vector), y is the actual value of ground truth, and $\hat{y} = F(x)$ is the predicted output on training set with size N. At each stage m, $1 \leq m \leq M$, a weak model F_m is to fit residual $h_m(x)$ to y-$F_{m-1}(x)$ by gradient boosting. That is, each F_m attempts to correct the errors of its predecessor model F_{m-1}. Therefore, the ensemble-learning model can be described as the optimization problem as follows.

$$\min_F \sum_{j=0}^{N} Lf\left(y_j, F_{m-1}(x_j) + h_m(x_j; \phi_m)\right) + \Omega(h_m)$$

$$F_m(X) = F_{m-1}(X) + h_m(X; \phi_m) \quad \text{(s.t)}$$

$$F_M(X) = \sum_{m=1}^{M} h_m(X; \phi_m)$$

Here, $x_j \in X$ is X^d and $y_j \in Y$ is Y^d respectively in Definition 4; ϕ_m is parameter of weak model F_m in m^{th} iteration, and F_M is the final ensemble model. Besides the loss function Lf, optimization goal includes $\Omega(h_m(x)) \sim (F_m\_\text{depth}, F_m\_\text{shrinkage})$, which is the regularization part restricted by depth and shrinkage rate of base trees. For such regularization, we refer the concepts in XGBoost [30]: tree depth controls model complexity (i.e., the degree of model can fit); shrinkage rate is a small extent to slow down the learning for a new base tree.

Therefore, prediction model can be trained as the procedure in Table 2 with multiple algorithmic parameters: vector temporal dimensionality θ, vector spatial dimensionality $|L|$, size of upstream dependent locations (i.e., spatial neighbour number) η, tree size (i.e., iterative number), maximal tree depth p, tree shrinkage r, and training set size N (i.e., size of historical days).

When the algorithm is completed, a regression model would be trained on historical data of previous N days. To initiate the model, $F_0(X)$ is generated by constant values in line 1 for any $Y^d \in Y$. After computing residuals in the lines 3–5, a base tree $h_m(X)$ as a weak model is found to fit X^d with those residuals in iterations. To find an approximation that minimizes average value of loss function Lf and regularization Ω, weight γ_m is achieved in line 6. Here, the regularization of base trees is controlled by parameters p and r. The model then incrementally expands itself by adding new weighted tree $\gamma_m h_m(X)$ as line 8. At last, the final model $F_M(X)$ as an ensemble of M base trees (i.e., after M iterations) is returned like line 10.

Through the trained model F_M, according to X^0 of feature V^0 in the current day, daily trends at any toll station in the next day $d \in D$ (i.e., D is prediction range) can be found from corresponding predictive Y^d. After these traffic trends are predicted, potential traffic hotspots detection is triggered to execute as the procedure in Fig. 2. For any $d \in D$, there are multiple on-premise top-K queries to find toll stations as hotspots after sorting traffic volumes in descending order. The time complexity of hotspots detection is $O(|L| * \log(|L|) * |D|)$ when Quick Sorting algorithm is adopted for sorting here.

Evaluation

Settings

In the project mentioned in Section 2.1, our method is evaluated by extensive experiments and case studies. Five Acer AR580 F2 rack servers via Citrix XenServer 6.2 are utilized to build a private Cloud, each of which own 8 processors (Intel Xeon E5–4607 2.20GHz), 64 GB RAM and 80 TB storage. To maintain raw toll data and aggregative traffic volume data of input data layer as Fig. 1, three virtual machines of the Cloud form a HBase 1.6.0 cluster, each of which owns 4 cores CPU, 22 GB RAM and 700 GB storage. In the practical scene of Henan highway, toll data from 343 toll stations (i.e., $|L| = 343$) would generate 1.5 million records a day. Another one of virtual machines (4 cores CPU, 8 GB RAM and 200 GB storage installing CentOS 6.6 x86_64 operating system) is used to install MySQL 5.6.17 as the relational database for both business profiles (station, section and highway line) and external data (i.e., calendar and meteorology data). The modules of feature modeling, hotspots detection and application are

Table 2 Model training to predict daily traffic trends

Algorithm: *regression model training through GBRT*
Input: feature vectors $\{V^d(X^d, Y^d)\}_{d=1-N}^{0}$ in previous N days (i.e., training size N), vector temporal dimensionality θ, vector spatial dimensionality $\|L\|$, size of upstream dependent locations η, tree size M, maximal tree depth p, and tree shrinkage r.
Output: ensemble model F_M for traffic trends prediction in next D days.
1. initialize a model with constant value γ : $F_0(X) = \arg\min_\gamma \sum_{d=1-N}^{0} Lf(Y^d, \gamma)$;
2. for m=1 to M
3. for d=(1-N) to 0
4. compute residuals: $r_d{}^m = Y^d - F(X^d)$;
5. endfor
6. find a base tree $h_m(X)$ to fit those residuals by training data $\{(X^d, r_d{}^m)\}_{d=1-N}^{0}$;
7. get weight γ_m by one-dimensional optimization: $\gamma_m = \arg\min_\gamma \sum_{d=1-N}^{0} Lf(Y^d, F_{m-1}(X^d) + \gamma h_m(X^d)) + \Omega(h_m)$;
8. update the model: $F_m(X) = F_{m-1}(X) + \gamma_m h_m(X)$;
9. endfor
10. return $F_M(X)$

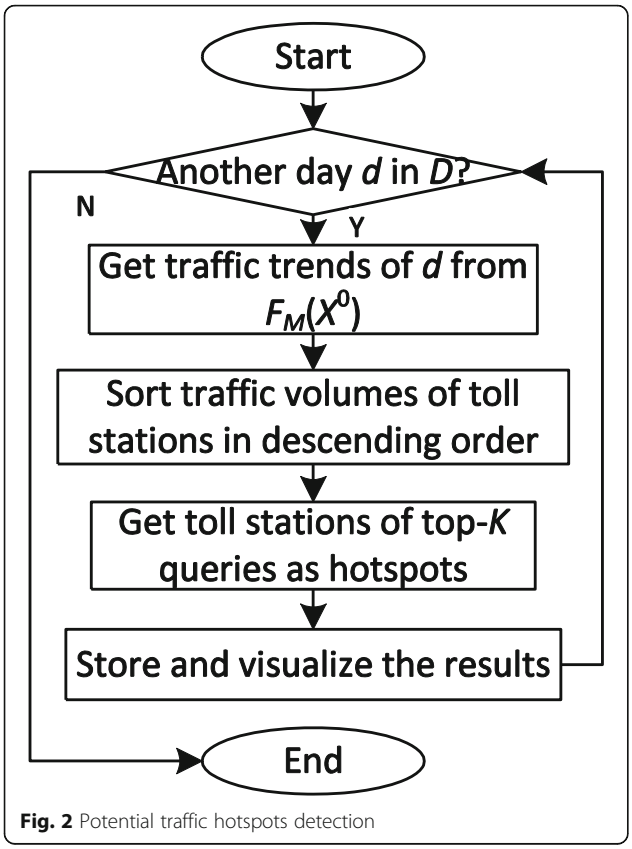

Fig. 2 Potential traffic hotspots detection

also implemented on that machine with Oracle JDK 1.7.0, an application server Apache Tomcat, and an open source machine learning toolkit scikit-learn 0.21.3.

As a routine business analysis on heterogeneous historical data of recent 120 days (i.e., $N = 120$), our detection method is triggered to execute at 12:00 a.m. each day and get the results for next 30 days (i.e., $D = 30$). More than 300 toll station is involved in this data (i.e., $|L| \sim 300$). All those results would be written to a dedicated table of HBase. According to the conclusion in our previous work [4], vector parameters $\theta = 10$ and $\eta = 4$.

To evaluate prediction effects, three metrics are adopted as follows.

Definition 5: Prediction metrics. Three terms of predictive error are used for experimental evaluation. The first is absolute percentage error (abbr. APE) defined as Eq. 1; the second is mean absolute percentage error (abbr. MAPE) as Eq. 2; the third is median absolute percentage error (abbr. MDAPE) as Eq. 3. Here, at a toll station in a given day, x_i and x_i' are respectively the factual value and the corresponding predictive one, where $i \in \mathbb{N}$ is the sequential index. APE is discrete values of any pair of x_i and x_i'; MAPE and MDAPE are aggregative statistics of those values respectively.

$$\mathrm{APE_i} = \frac{|x_i - x_i'|}{x_i} * 100\% \tag{1}$$

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|x_i - x_i'|}{x_i} * 100\% \tag{2}$$

$$\mathrm{MDAPE} = \mathrm{median}\ (\mathrm{APE}). \tag{3}$$

Then, our method is evaluated in three parts: algorithmic parameters tuning, trends prediction, and hotspots detection case studies.

Parameters tuning and traffic trends prediction

As the design of Section 3, a traffic trends prediction model would be trained with proper algorithmic parameters: tree size (i.e., iterative number) M, maximal tree depth d, and tree shrinkage (i.e., learning rate) r. Here, M, $d \in \mathbb{N}$, and $r \in \mathbb{R}^+$. In practice, the model has to be retrained periodically (e.g., once a day, a week, or a month) to fit recent trends better. Depth parameter d has upper bound 6 [31], and is set median 3 constantly here to trade-off model complexity and tree-structure split efficiency. A following experiment is designed first to find proper parameters of the model.

Experiment 1: parameters tuning

The traffic volumes data from April to July of 2017 is employed as training set, and that of August 2017 is adopted as validation set for the prediction model. Besides such business data, meteorology data and calendar data of corresponding days are imported to build feature vectors of Definition 4 through Hadoop MapReduce. After setting $\theta = 10$, $\eta = 4$, $D = 30$, and $d = 3$ as mentioned before, we want to find proper pairs of M and r for the model. Parameter M is set as 500, 1000, 2000, 2500, 3000 and 5000 respectively, and r is set as 0.005, 0.01, 0.05, 0.1 and 0.5 then. Under each combination of M and r on the training set, the executive time of a prediction and metric MAPE between predictive values and ground truth would be counted in average. A specific location *Zhengzhou South*, one of toll stations with heaviest traffic in Henan province, is chosen for evaluation.

The results are illustrated as Fig. 3. We found executive time of our method rises when M grows from Fig. 3 (a). Such longer latency appears because larger M implies more iterations, more base trees, and much executive time. Contrarily, the latency is insensitive to r: under the same M, executive times here present few differences when r grows ten-fold. With a small depth $d = 3$ regardless of r, all base trees split quickly. From Fig. 3 (b), when M grows, metric MAPE with different r drops distinctly, keeps steady, and even grows then. Overfitting appears when M is too large. Referring to the latency in Fig. 3 (a), the combination of $M = 3000$ and $r =$

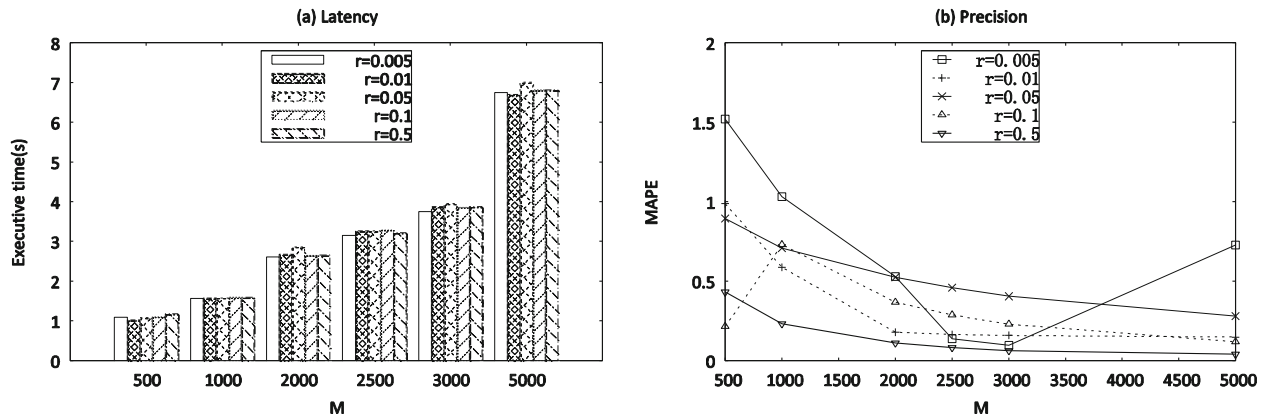

Fig. 3 Parameters tuning

0.5 is appropriate because precision is in a steadily high level with the latency less than 4 s. It just reflects the trade-off between efficiency and effectiveness for trends prediction.

Then, another experiment is conducted to show the precision of traffic trends prediction through the trained model (i.e., $M = 3000$, $d = 3$, $r = 0.5$, $N = 120$, and $D = 30$). In order to quantitatively evaluate predictive effects of our model (abbr., GBRT), three variants and two other models are also implemented for comparison in hotspot detection module of Fig. 1. According to Definition 4, GBRT_w does not consider meteorological dimension, GBRT_d abandons calendric dimension, and GBRT_wd is a variant without both dimensions of meteorology and calendar. Moreover, on the same training set and validation set as the experiment above, ARIMA (Auto-Regressive Integrated Moving Average) as a linear model and SVR (Support-Vector Regression) as a non-linear model are employed as additional counterparts. For the former one, the model [6] is adopted as ARIMA (p, d, q), where algorithmic parameter p is autoregressive size, q is number of moving average size, and d is steps number for stationary differences. This model is tuned with the optimal parameters $p = 2$, $d = 1$ and $q = 3$. For the latter one, the model [32] is adopted as SVR (*kernel*, C, ε), where algorithmic parameter *kernel* is kernel function, C is penalty of errors, and ε specifies epsilon-tube distance. This model is tuned with the parameters *kernel* = 'rbf', $C = 500$, and $\varepsilon = 0.8$ accordingly.

Experiment 2: predictive effects

The data of September 2017 is employed as test set for our trained model and five counterparts above during traffic trends prediction. Through each model in a prediction for all locations in the highway network, executive latencies are counted, and metrics APE, MAPE and MDAPE at the toll station *Zhengzhou South* are calculated after the finish of prediction procedures.

Predictive effects are illustrated as Fig. 4 and Table 3: the figure shows predictive effects in 30 days; the table presents executive latency and aggregative precision by the metrics of MAPE and MDAPE. We found our method performs best in time consumption with relatively low predictive errors. (1) All six models can approximate daily traffic trends in the whole month, and GBRT with its variants drop predictive errors about 10% than the other twos. ARIMA and SVR perform a similar order of magnitude in predictive precision. As Fig. 4(a),

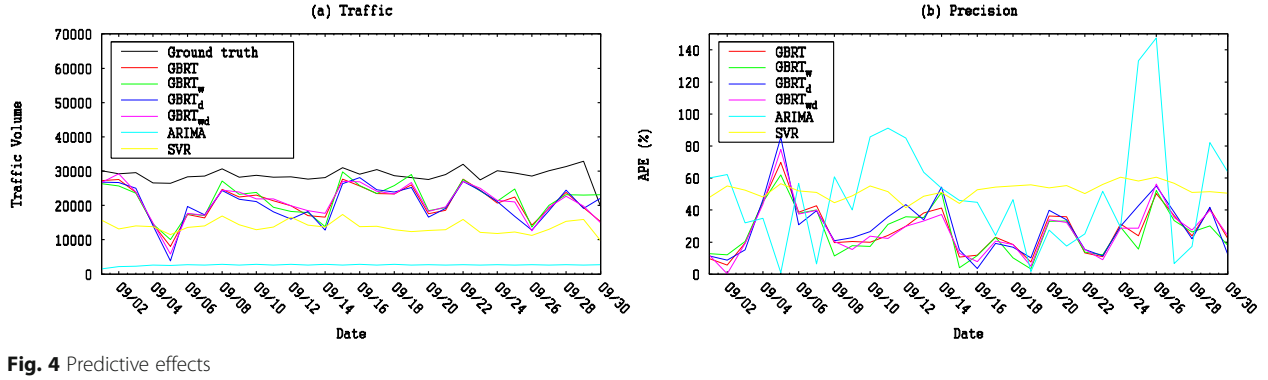

Fig. 4 Predictive effects

Table 3 Efficiency and precision

	GBRT	$GBRT_w$	$GBRT_d$	$GBRT_{wd}$	ARIMA	SVR
MAPE(%)	27.38	25.94	29.43	26.93	49.930	52.34
MDAPE(%)	23.89	24.61	28.21	25.91	46.21	52.27
Latency(s)	9.34	8.82	8.66	7.92	41.26	50.18

traffic bursting and shrinkage can be presented by all those models but with more fluctuation. As Fig. 4 (b) and Table 3, such predictive effects are acceptable. In Table 3, GBRT and its variants reduce MAPE distinctly than SVR or ARIMA, which comes from multiple dimensions on heterogeneous data for prediction. On single type data of traffic volumes, ARIMA merely uncovers one dimensional time-series and SVR gets inexplicable temporal and spatial feature jointly. Without heterogeneous data both models would lose much information about spatio-temporal business correlation. (2) GBRT and its variants have the evident advantages in executive latency. They cost less than 10 s due to fast convergence of the algorithm as Table 2, while the counterparts require almost one minute correspondingly. Moreover, during a prediction, GBRT and its variants can achieve predictive values of all locations, but others work only for a single one. It makes our method more practical and efficient in domain analytics. (3) Among the variants, $GBRT_w$ and GBRT have better precision than others. $GBRT_w$ performs a little better in MAPE because no extreme weather in days of test set leads more efficient calculation. In executive performance perspective, more dimensions a model considers, much time it would consume. It is the reason that GBRT requires longer latency than its variants. However, the latency increment is no more than 1.5 s and is worthy to get better predictive precision. In summary, our method performs well in performance and holds a high precision in traffic trends prediction.

Case study of potential hotspots detection

We then evaluate potential hotspots detection in two case studies via our system mentioned in Section 2.1. One is an overview effects, and the other is statistics in multiple perspectives.

Case 1

Figure 5 is parts of the index page in our system, and two maps are included. The left is an administrative map of *Henan* province, where heat degree of cities is presented in colors. The right is a map of provincial city *Zhengzhou* with surrounding toll stations, where heat degree of stations is presented in colored bubbles. Data of current day is employed and the potential hotspot stations of the next day are available here.

The heat degrees of both maps in Fig. 5 are evaluated by the exit traffic volume of Definition 4. Heat degree of a city in the left map is the sum of traffic volumes at the toll stations located in that city. According to quantities of traffic volumes, the heats are presented in different

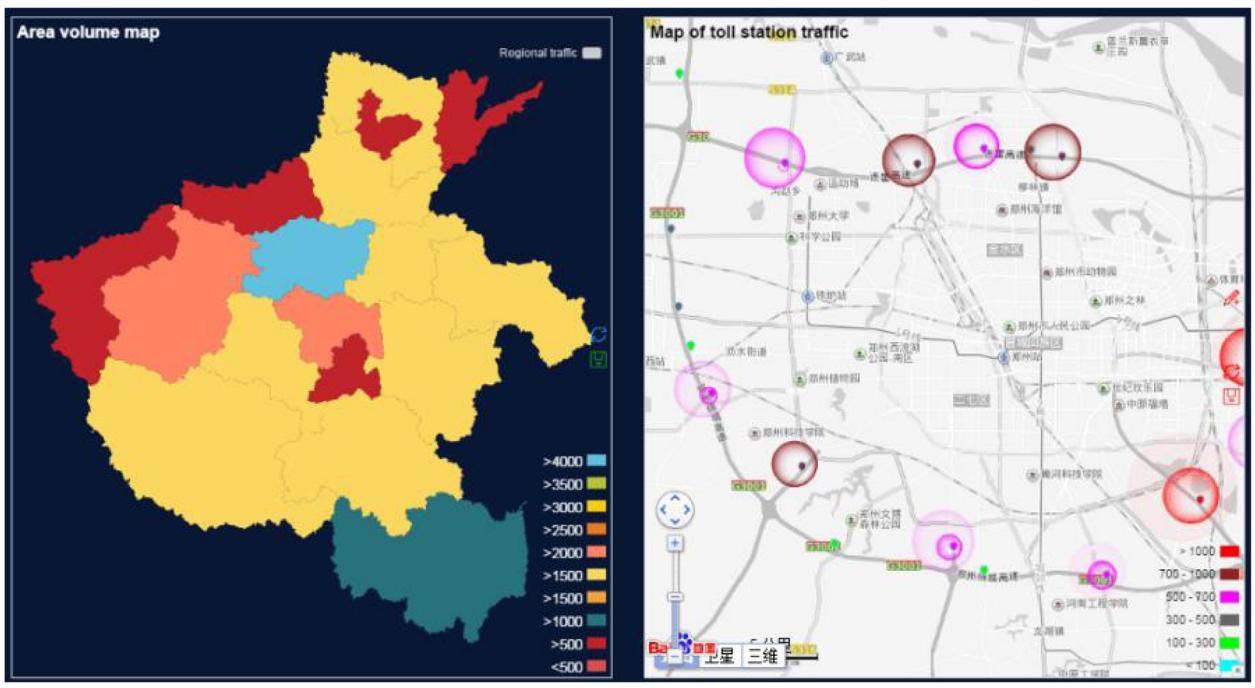

Fig. 5 Potential hotspots overview

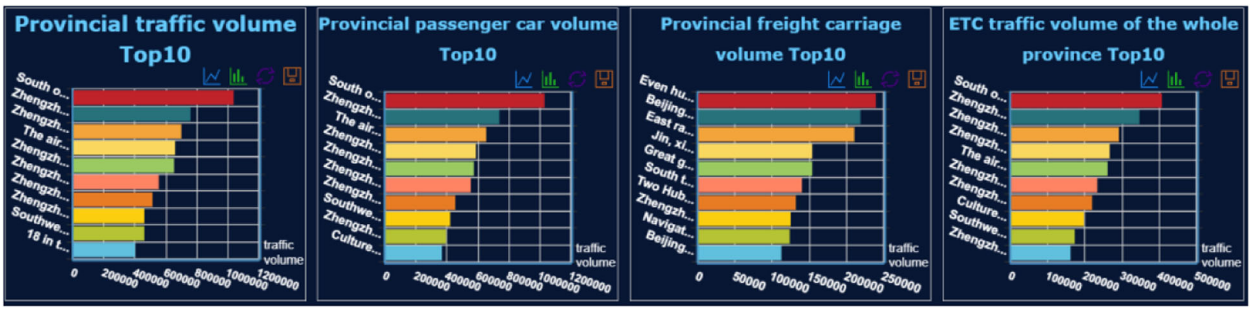

Fig. 6 Potential hotspots in multiple perspectives

colors. We found the central north city Zhengzhou, the provincial city of Henan, has the largest amount in blue color. In an interactive fashion, after clicking a city in the left map, we can find a detailed map of that city in the right. City Zhengzhou is shown by default as the right map in Fig. 5, where ten more toll stations are located in its peripheral. The potential traffic hotspots are displayed here in colored bubbles. The darker the red color of a toll station is, the hotter of this station would be. We can directly find future trends that east of Zhengzhou would be busy in the next day, because two of toll stations in east are top level of potential hotspots. On daily traffic trends at toll stations surrounding Zhengzhou, such result of potential hotspots is from a top-K query where $K = 10$. In interactive maps with visualization, our method for potential traffic hotspots is directly accessible for business technicians.

Case 2

Figure 6 is parts of daily situation in a statistical page. In four perspectives, toll stations of potential traffic hotspots are presented in bar charts. Historical data of June–August 2017 is employed here.

As parts of a historical statistics page, potential traffic hotspots for 1st September 2017 are showed in four perspectives as Fig. 6. Each of them is from a respective top-K query where $K = 10$. The first one is from the daily traffic trends on toll data of all types of vehicles at stations in Henan province. The second and the third are on data of passenger cars and freight carriages respectively. The fourth one is on data of ETC vehicles. That is, with queries of traffic trends on specific data, potential hotspots of customized requirements would be built by our method with visualization, without much professional programing knowledge. Moreover, compared with ground truth data in the next day, the results here fit factual values well. It proves the feasibility and effectiveness of our method in practice.

Conclusion

In this paper, a novel ensemble-learning method is proposed for potential traffic hotspots detection in highway domain. On heterogeneous spatio-temporal data, such as toll data, meteorological data and calendric data, traffic volume feature is built integrally considering characteristics about time, space, weather, and date factors at stations of highway network. Through gradient boosting regression technology, low executive latency is kept and meanwhile predictive error can be improved distinctly than traditional models for traffic trends prediction. In extensive experiments and case studies through our method on real data of one Chinese province, the time consumption is hold in less than 5 s with nearly 10% precision improvement. Potential hotspots detection of multiple perspectives is done by interactive online map with visualization.

Due to fluctuant predictive precision on some special holidays, daily traffic trends in those periods have to be improved next. In our future work, such precision may perform better with larger training set and dedicated feature of fine-grained attributes accordingly.

Acknowledgments

Thanks to my son baby Jiayun Ding and wife Jingyi Qin, the housework for you with happiness pushes me to perform my research in a more efficient manner.

Authors' contributions

Weilong Ding has written this paper and done the research which supports it. All the authors have collaborated in conception, research and design, and approved the final manuscript.

Funding

This work was supported by National Natural Science Foundation of China (No. 61702014, 61702491), Beijing Municipal Natural Science Foundation (No. 4192020), Top Young Innovative Talents of North China University of Technology (No. XN018022), and "Yuyou" Talents of North China University of Technology.

Availability of data and materials

The datasets of this article to support the findings have not been made available because they were supplied by local officials of *Henan Transport Department* with certain confidentiality level.

Competing interests
The authors declare that there is no conflict of interest regarding the publication of this paper.

Author details
[1]School of Information Science and Technology, North China University of Technology, Beijing 100144, China. [2]Beijing Key Laboratory on Integration and Analysis of Large-scale Stream Data, Beijing 100144, China. [3]Big Data Center, State Grid Corporation of China, Beijing 100031, China. [4]China Electric Power Research Institute, Beijing 100192, China. [5]Institute of Microelectronics, Chinese Academy of Sciences, Beijing 100029, China.

Received: 17 December 2019 Accepted: 7 April 2020
Published online: 11 May 2020

References

1. Li Z, Li Y, Li L (2014) A comparison of Detrending models and multi-regime models for traffic flow prediction. IEEE Intell Transp Syst Mag 6(4):34–44
2. Zhu F, Lv Y, Chen Y, Wang X, Xiong G, Wang F (2019) Parallel Transportation Systems: Toward IoT-Enabled Smart Urban Traffic Control and Management. In: IEEE Transactions on Intelligent Transportation Systems, pp 1–9
3. Yi X, Duan Z, Li T, Li T, Zhang J , Zheng Y (2019). CityTraffic: Modeling Citywide Traffic via Neural Memorization and Generalization Approach. In 28th ACM International Conference on Information and Knowledge Management (CIKM 2019), Beijing, China
4. Ding W, Wang X, Zhao Z (2020) CO-STAR: a collaborative prediction service for short-term trends on continuous spatio-temporal data. Futur Gener Comput Syst 102:481–493
5. Gao H, Xu Y, Yin Y, Zhang W, Li R, Wang X (2019) Context-aware QoS prediction with neural collaborative filtering for internet-of-things services. IEEE Internet Things J:1
6. Ding W, Zhao Z (2018) DS-harmonizer: a harmonization service on Spatio-temporal data stream in edge computing environment. Wirel Commun Mob Comput 2018:12
7. Ding W and Cao Y. A Data Cleaning Method on Massive Spatio-Temporal Data. Advances in Services Computing: 10th Asia-Pacific Services Computing Conference, APSCC 2016, Zhangjiajie, China, November 16–18, 2016, Proceedings, Wang G, Han Y and Martínez Pérez G (eds). Springer International Publishing, 2016, pp. 173–182. Springer, Cham
8. Ghesmoune M, Lebbah M, Azzag H (2016) State-of-the-art on clustering data streams. Big Data Analytics 1(13):1–27
9. Zheng Y, Capra L, Wolfson O, Yang H (2014) Urban computing: concepts, methodologies, and applications. ACM Trans Intell Syst Technol 5(3):1–55
10. Targino AC, Gibson MD, Krecl P, Rodrigues MVC, dos Santos MM, de Paula Corrêa M (2016) Hotspots of black carbon and PM2.5 in an urban area and relationships to traffic characteristics. Environ Pollut 218:475–486
11. Xie K, Ozbay K, Kurkcu A, Yang H (2017) Analysis of traffic crashes involving pedestrians using big data: investigation of contributing factors and identification of hotspots. Risk Anal 37(8):1459–1476
12. Puertas E, Fernández J, Morales-Botello M, Aliane N (2013) Detection and visualization of potential traffic hotspots in urban environments. In: 13th International Conference on ITS Telecommunications (ITST2013), Tampere, Finland, pp 85–89
13. Solé-Ribalta A, Gómez S, Arenas A (2016) A model to identify urban traffic congestion hotspots in complex networks. R Soc Open Sci 3(10):160098
14. Rüegg T, Wittneben A (2018) Resource Allocation for QF VMIMO Receive Cooperation in Urban Traffic Hotspots. In: 26th European Signal Processing Conference (EUSIPCO2018), Rome, Italy, pp 1502–1506
15. Jaziri A, Nasri R, Chahed T (2016) Offloading traffic hotspots using moving small cells. In IEEE International Conference on Communications (ICC2016), Kuala Lumpur, Malaysia, pp 1–6
16. Lingras P, Sharma SC, Osborne P, Kalyar I (2000) Traffic volume time-series analysis according to the type of road use. Computer-Aided Civil and Infrastructure Engineering 15(5):365–373
17. Min W, Wynter L, Amemiya Y (2011) Real-time road traffic prediction with spatio-temporal correlations. Transportation Res C Emerg Technol 19(4):606–616
18. Wang Y, Papageorgiou M (2005) Real-time freeway traffic state estimation based on extended Kalman filter: a general approach. Transp Res B Methodol 39(2):141–167
19. Li D-M, Liu B (2014) Modeling and prediction of highway traffic flow based on wavelet neural network. In: 2014 International Conference on Machine Learning and Cybernetics, Lanzhou, Gansu, China, pp 675–679
20. Abdulhai B, Porwal H, Recker W (2002) Short-term traffic flow prediction using Neuro-genetic algorithms. J Intell Transp Syst 7(1):3–41
21. Chan KY, Dillon T, Chang E, Singh J (2013) Prediction of short-term traffic variables using intelligent swarm-based neural networks. IEEE Trans Control Syst Technol 21(1):263–274
22. Wosyka J, Přibyl P (2012) Real-time travel time estimation on highways using loop detector data and license plate recognition. In: ELEKTRO 2012, pp 391–394
23. Meng C, Yi X, Su L, Gao J and Zheng Y (2017). City-wide Traffic Volume Inference with Loop Detector Data and Taxi Trajectories. In Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Redondo Beach, CA, USA, pp. 1–10
24. Chen C, Li K, Teo S G, Zou X and Wang K (2019). Gated Residual Recurrent Graph Neural Networks for Traffic Prediction. In 31th AAAI Conference on Artificial Intelligence (AAAI 2019), Honolulu, Hawaii, USA
25. He Z, Chow C, Zhang J (2019) STCNN: A Spatio-Temporal Convolutional Neural Network for Long-Term Traffic Prediction. In: 2019 20th IEEE International Conference on Mobile Data Management (MDM), pp 226–233
26. Guo S, Lin Y, Feng N, Song C and Wan H (2019). Attention Based Spatial-Temporal Graph Convolutional Networks for Traffic Flow Forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence, pp. 922–929
27. Gao H, Duan Y, Shao L, Sun X (2019) Transformation-based processing of typed resources for multimedia sources in the IoT environment. Wireless Networks
28. Xia Y, Wang X and Ding W (2018). A Data Cleaning Service on Massive Spatio-Temporal Data in Highway Domain. In Service-Oriented Computing – ICSOC 2018 Workshops, Hangzhou, China, pp. 229–240
29. Administration CM (2010) Meteorological industry standards of China: grade of weather conditions for freeway transportation (QXT111–2010, in Chinese)
30. Chen T and Guestrin C (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, pp. 785–794
31. Steadman M (2014) Gradient Boosted Regression Trees. DataroRot, p 2019
32. Ding W, Zhao Z, Wang J and Li H (In press). Task allocation in hybrid big data analytics for urban IoT applications. *ACM/IMS Transactions on Data Science*

Publisher's Note