An Unlicensed Taxi Identification Model Based on Big Data Analysis

Wei Yuan, Pan Deng, Tarik Taleb, *Senior Member, IEEE*, Jiafu Wan, *Member, IEEE*, and Chaofan Bi

***Abstract*—Social networks and mobile networks are exposing human beings to a big data era. With the support of big data analytics, conventional intelligent transportation systems (ITS) are gradually changing into data-driven ITS (D^2ITS). Along with traffic growth, D^2ITS need to solve more real-life problems, including the issue of unlicensed taxis and their identification, which potentially disrupts the taxi business sector and endangers society safety. As a remedy to this issue, a smart model is proposed in this paper to identify unlicensed taxis. The proposed model consists of two submodel components, namely, candidate selection model and candidate refined model. The former is used to screen out a coarse-grained suspected unlicensed taxi candidate list. The list is taken as an input for the candidate refined model, which is based on machine learning to get a fine-grained list of suspected unlicensed taxis. The proposed model is evaluated using real-life data, and the obtained results are encouraging, demonstrating its efficiency and accuracy in identifying unlicensed taxis, helping governments to better regulate the traffic operation and reduce associated costs.**

***Index Terms*—Big data, intelligent transportation systems, machine learning, data-driven ITS, unlicensed taxi.**

I. Introduction

ALONG with the rapid growth of urban economies, cities are expanding rapidly and urban populations are constantly increasing. At the same time, urban population floating range are largely increasing and land resources are gradually becoming tense, which results in serious traffic congestions, traffic safety problems and environmental pollution. All these issues have become public concerns. At this point, Intelligent Transportation Systems (ITS) appear as a solution to some concerns [1], [33], [34]. In the past two decades, experts all over the world have made lots of research advancing the concept of ITS to ultimately better manage urban traffic and improve the quality of life of people. Some of the conducted research has largely contributed to ensuring traffic safety [34] and better use of road resources [2], [35], [36]. Although the application of ITS solves some urban transportation problems, to a certain extent, the constant increase in current transportation data does not meet up with the demand for data usage, which is crucial for transportation control. To fully utilize data, the concept of Data-Driven ITS (D^2ITS) has emerged [3]. D^2ITS has evolved from the conventional ITS but with some differences. First, the conventional ITS systems depend mainly on historical data and people's real life experience [37]–[39] whereas D^2ITS systems tend to use real-time data. Second, in conventional ITS systems, the source of data used is generally limited whilst D^2ITS can make good use of all types of data, including vehicle information, sense coil and video images. Along with the ongoing advances in big data analytic techniques, D^2ITS has become increasingly powerful providing new solutions to several problems, deemed unsolvable before.

Big data technologies could quickly extract valuable information from large amount and various types of data [4]. Big data are characterized by four typical features: 1) huge volume exceeding TB level to PB level; 2) various data types including videos, photos, location information, and sensor data; 3) low value density and high application value (e.g., within a video used to monitor road conditions, there may be only one or two minutes useful clips); and 4) fast processing speed. In the transportation field, there are four types of data that are frequently used: 1) sensor data (e.g., position data, temperature data, pressure data, images, and RFID); 2) system data (e.g., logs and equipment records); 3) service data (e.g., charging information, internet services, and some other information); and 4) application data (e.g., manufacturers, energy, and transportation). Transportation-relevant data consist indeed of various types and enormous volume which, in turn, need fast processing speed.

One type of transportation data is the pass-records data collected from the Electronic Traffic Bayonet Device (ETBD), which has high density and significant value for data mining and analysis. ETBD [5] integrates advanced optical technologies, image processing technologies and pattern recognition

Manuscript received November 8, 2014; revised April 1, 2015 and July 16, 2015; accepted October 6, 2015. Date of publication November 24, 2015; date of current version May 26, 2016. This work was supported in part by the National Natural Science Foundation of China under Grants 61100066, 61472283, 61572220, and 61262013; by the Fok Ying-Tong Education Foundation of China under Grant 142006; by the Fundamental Research Funds for the Central Universities under Grants 2100219043, 1600219246, and x2jq-D2154120); and by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry. The Associate Editor for this paper was W.-H. Lin. *(Corresponding author: Jiafu Wan.)*

W. Yuan and C. Bi are with the Institute of Software, Chinese Academy of Sciences, Beijing 100190, China (e-mail: futureyuan628@gmail.com; bichaofan@gmail.com).

P. Deng is with the Institute of Software, Chinese Academy of Sciences, Beijing 100190, China, with Guiyang Academy of Information Technology, Guiyang 550000, China, and also with Guiyang Technology Bureau, Guiyang 550081, China (e-mail: dengpan@iscas.ac.cn).

T. Taleb is with the School of Electrical Engineering, Aalto University, 02150 Espoo, Finland (e-mail: talebtarik@ieee.org).

J. Wan is with the School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou 510640, China (e-mail: jiafuwan_76@163.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2015.2498180

technologies. It can take photos of each passing vehicle and automatically identifies the license plate, brand, type, speed and some other information relevant to the vehicle. It then stores the information into a database.

The pass-records data have very high values in the field of D^{2}ITS. They can be used for traffic volume prediction, traffic trajectory tracking, and traffic behavior analysis. These data are crucial for the supervision and administration of traffic violations and tracking of criminals. In this vein, unlicensed taxis function like a proper taxi carrying passengers between different locations but without transport business licenses and tax registration certificates [6]. In other words, unlicensed taxis are legally not allowed to carry passengers. Impact of unlicensed taxis can be mainly summarized into four aspects [42], [43].

1. Unlicensed taxis seriously disrupt traffic operations and induce safety issues. Indeed, most unlicensed taxi drivers lack necessary security awareness. They tend to more likely ignore traffic regulations.
2. In case of unlicensed taxis, vehicle safety cannot be always guaranteed. Indeed, checkups of vehicle conditions are not carried out on a regular basis.
3. Given the above, passengers' safety cannot be guaranteed either. Furthermore, because of unlicensed taxi drivers' complicated background and relatively low level education, passengers may become subject to drivers' inappropriate behavior.
4. In case of a traffic accident, passengers' legal rights cannot be guaranteed as unlicensed taxis are not registered at the transportation department. Once an accident happens, it becomes difficult to trace back the cause.

Conventionally, the police department takes the following actions to identify unlicensed taxis and punishes the drivers.

1. Find out suspicious unlicensed taxis through manually watching amounts of videos in the "City Online Monitoring System";
2. Collect the drivers' on-site evidence by entrapment;
3. Make appropriate punishment on the suspects.

This process not only wastes a lot of manpower and time, but also results in having many unlicensed taxis go undetected. It is too passive, and a more proactive method is proposed in this paper.

The objective of this paper is to intelligently identify such unlicensed taxis conducting analysis and data mining on the pass-records data. The solution proposed in this paper will hopefully put an end to the unlicensed taxi phenomenon, saving manpower and time of police officers that, otherwise, have to stay for long durations monitoring for suspicious unlicensed taxis. Although some D^{2}ITS technologies, such as image processing and cloud computing have set up a solid foundation for data generation, data acquisition, and data storage [7], review of existing literature does not reveal any approach suitable for unlicensed taxi identification. In this regard, this paper proposes an algorithm based on big data analysis, integrating Support Vector Machine (SVM) [8], [47] learning, whereby Hadoop [9] and Map/Reduce [10] technologies are adopted as well to generate data. Compared with traditional approaches whereby unlicensed taxis are identified by patrolling police, the proposed method achieves its design goals with high accuracy and efficiency. It shortens the unlicensed taxi identification cycle and saves important amount of manpower, eventually yielding important cost savings in the long run.

The paper is organized in the following fashion. Section II discusses some research work relevant to D^{2}ITS. Section III portrays the approach adopted for data acquisition, data preprocessing, and data modeling. Our scheme proposed for unlicensed taxi identification is detailed in Section IV. Some test results are presented and discussed in Section V. Finally, the paper concludes in Section VI.

II. Related Work

Being an essential field of transportation engineering, ITS consists of six fundamental components; namely advanced transportation management systems (ATMS), advanced traveler information systems (ATIS), advanced vehicle control systems (AVCS), business vehicle management (BVM), advanced public transportation systems (APTS), and advanced urban transportation systems (AUTS). As discussed in [3], whether these six components can be fully realized depends on how much data can be collected and processed into useful information. As social networks and mobile networks are developing at a fast speed, data start to overflow, evolving into the age of big data. With large amount of ITS-relevant data, ITS is gradually changing into the so-called D^{2}ITS. There are two major differences between ITS and D^{2}ITS. First, conventional ITS systems depend mainly on human experiences and historical data, while D^{2}ITS systems are based on real time data [3]. Second, data used by conventional ITS systems are collected from limited sources, whilst D^{2}ITS systems collect data from multiple sources. Based on big data, D^{2}ITS systems exhibit better performance. They are straightforward and practical in real life.

A main component of D^{2}ITS is the Vision-Driven ITS, which employs vision-based devices (e.g., cameras) to collect data for doing research and developing applications. In [11] and [12], a Pedestrian Tracking System is developed to detect the crossing characteristics of pedestrians at intersections. Background subtraction technique, tracked through an inherent cost characteristic function in conjunction with an $\alpha-\beta$ filter, is used. This advanced system introduces an automated computer-vision based approach for collecting pedestrian parameters in real-time. A similar approach dedicated for vehicle detection and tracking is proposed in [13]. In [14], another D^{2}ITS system, called DRIVE Net, is developed for data sharing, visualization, modeling and analysis and that is using data extracted from different sources. The basic intention beneath this DRIVE Net platform is to offer a standard tool to incorporate more datasets from different fields and at the same time, and that is for real-time decision making. In [15], a Web-based truck performance measurement system is developed processing commercial GPS

data and quantifying truck travel characteristics and performance between zones. In [16], a Markov chain based Bayesian decision tree algorithm is devised to extract passengers' origin information from recorded transit smart card systems, and that is for transit system planning and route optimization. In [17], pedestrian movement monitoring is conducted using static Bluetooth sensors, enabling an automated and cost-effective approach to acquire pedestrian data.

The latest developments in ITS pay more attention to proactive control as opposed to conventional passive control and management [18]. This has yielded Learning-Driven ITS which is based on different schemes. Among all of them, real-time traffic prediction schemes, such as those used for the prediction of travel time [37], [38], vehicle passage time [39], and hazardous locations are vital, and have been proven important for the support of efficient communications for highly mobile users in mobile networks [40], [41]. In [19], a state-space neural network is proposed to conduct short-term freeway travel time prediction. A real-time Kalman filter model to predict travel time by combining historical data with real-time measurements is proposed in [20]. Another cluster-based algorithm to identify the collision hotspots along a freeway is proposed in [21].

Thanks to big data analytic techniques, some problems which were impossible to solve before have solutions nowadays. For example, in populated countries such as China, a potential number of people tend to use fake vehicle licenses to avoid being fined for illegal travel behavior. In the past, data collection infrastructure was not available. Stopping the phenomenon of fake vehicle licenses therefore required important human labor, which was not always effective [22]. Nowadays, traffic flow collection system can be setup to capture the pictures of vehicles and image processing technologies can be subsequently used to extract useful information [22]. With the abundance of data, the fake vehicle license phenomenon can be mitigated to some degree. In this paper, we also aim at solving another "popular" transportation issue, namely unlicensed taxis, and that is using big data analytic technologies.

III. Data Acquisition

A. Dataset Description

In this paper, we use vehicle trajectory, consisting of temporal and spatial characteristics of vehicle movements, to identify unlicensed taxis. The method adopted for getting vehicle trajectory is different from the conventional way which traditionally uses GPS [23]–[25], [46]. Indeed, the conventional way is relatively much more expensive as it requires all vehicles to be equipped with GPS devices. In this paper, vehicle trajectory is basically deducted from ETBD since vehicles have to continuously pass many ETBDs during the course of their motion. A vehicle passing record means a vehicle is captured by some ETBDs and at the same time one corresponding recognition dataset is generated. The dataset includes vehicle passing record ID, vehicle license, vehicle type, collection date, and ETBD ID. On average, there are almost 20 million records generated on a daily basis and each record may consist of 20 attributes. Given the big data nature of the dataset and in order to maintain acceptable data processing speeds, the dataset is filtered removing unnecessary data attributes and keeping only useful attributes as shown in Table I.

TABLE I
A Sample of Vehicle Passing Records

Vehicle License ID	Vehicle License	Vehicle Type	Collection Date	ETB D ID	Vehicle Speed
BIZ2014010 80718193521 32074111132	xxxZ373	02	2014-01-08 07:18:12	1011 017	31
BIZ2014011 20607116301 93881667126	xxx207F	02	2014-01-12 06:07:07	1011 265	26
BIZ2014012 70848046871 88926845481	xxxA823	02	2014-01-27 08:51:19	1011 198	45

TABLE II
ETBD Dataset Attributes

ETBD ID	Latitude	Longitude	ETBD Address
1011017	106.69364	26.58523	intersection between xx and xx
1011265	106.71517	26.58532	intersection between xx and xx
1011198	106.71539	26.60548	intersection between xx and xx

In Table I, vehicle license ID (referred to as nid throughout this paper) is the unique identification of vehicle passing record and is a 32 char string. Vehicle license (referred to as vLicense) is the license of the detected vehicle. Vehicle type (denoted by vType) indicates the type of vehicle and can be, amongst many types, police car, full-size car, or small car. Collection date (denoted by vDate) refers to when the corresponding record is generated. ETBD ID (denoted by etbdID) is the unique identifier of the corresponding ETBD device which captures the vehicle and generates the dataset. Vehicle speed (denoted as vSpeed) refers to the speed of the vehicle at the collection date time. Accordingly, one vehicle passing record can be formulated as follows:

$$r = (\text{nid}, v\text{License}, v\text{Type}, v\text{Date}, \text{etbdID}, v\text{Speed}). \quad (1)$$

ETBD devices are usually installed at fixed locations in the city and can capture all vehicles passing by one or more lanes. Another used dataset is the ETBD dataset which contains several location attributes, as shown in Table II.

In Table II, ETBD ID is the same as the one in Table I. Latitude, longitude and ETBD address indicate the relevant ETBD's location. This dataset can be formulated as follows:

$$\text{etbd} = (\text{etbdID}, \text{latitude}, \text{longitude}, \text{etbdAddress}). \quad (2)$$

Merging the two datasets based on the common field $etbdID$, a vehicle passing record can be formulated as follows, giving all information required for the analysis.

$$r' = (\text{nid}, v\text{License}, v\text{Type}, v\text{Date}, v\text{Speed}, \text{latitude}, \text{longitude}, \text{etbAddress}). \quad (3)$$

B. Data Preprocessing

In a single city, there could be several hundreds of thousands of vehicles on the move. They would generate 10 million records (i.e., roughly 30 GBytes) on a daily basis. Efficient preprocessing of this big data becomes then required. In this vein, and in addition to filtering out unnecessary fields from the original vehicle passing record, "bad data," caused by ETBD deviation, should be eliminated. Bad data can lead to biased results, resulting in wrong conclusions [26]. Typical bad data are those with key fields missing or invalid. There are still less than 1% bad data which need more intelligent algorithms to detect. After the elimination of these bad data and on all vehicle passing records, denoted by group $\mathrm{R} = \{r_1, r_2, \ldots, r_n\}$ whereby $r_i = (\mathrm{nid}_i, v\mathrm{License}_i, v\mathrm{Type}_i, c\mathrm{Date}_i, \mathrm{etbdID}_i, v\mathrm{Speed}_i)$ five kinds of statistics are extracted using distributed systems [27], [44], [45].

i. Counting the number of vehicle passing records in time τ

$$C_{(\mathrm{vpr},\tau)} = \sum (r_i | c\mathrm{Date}_i \in \tau). \tag{4}$$

ii. For every ETBD, counting the number of vehicle passing records in time τ

$$C_{(\mathrm{vpr},\mathrm{etbdID},\tau)} = \sum (r_i | c\mathrm{Date}_i \in \tau, \mathrm{etbdID}_i = \mathrm{etbdID}). \tag{5}$$

iii. For every vehicle, counting the number of vehicle passing records in time τ

$$C_{(\mathrm{vpr},v\mathrm{License},\tau)} = \sum (r_i | c\mathrm{Date}_i \in \tau, v\mathrm{License}_i = v\mathrm{License}). \tag{6}$$

iv. Counting the total number of vehicles in time τ

$$C_{(\mathrm{vehicle},\tau)} = \sum (r_i | c\mathrm{Date}_i \in \tau, \text{Group by } v\mathrm{License}). \tag{7}$$

v. Counting the number of ETBD devices crossed by a vehicle during a time interval τ

$$C_{(etbd,vehicle,\tau)} = \sum (r_i | c\mathrm{Date}_i \in \tau, v\mathrm{License}_i = v\mathrm{License}, \text{Group by etbd}). \tag{8}$$

vi. Counting the total speed for all vehicle-passing records in time τ

$$C_{(\mathrm{speed},\tau)} = \sum (r_i | v\mathrm{Speed}_i \in \tau). \tag{9}$$

C. Data Modeling

In this paper, we use hotspot, namely the distribution of the number of vehicle passing record, as the research object. The location of each ETBD is defined by its latitude and longitude on a map. All ETBDs are classified based on the number of vehicle-passing records using a K-means [28] clustering algorithm based on $C_{(\mathrm{vpr},\mathrm{etbdID},\tau)}$. All ETBDs are categorized into h grades, denoted by $\mathrm{ETBD\_heat}_i$, whereby $i \in [1, h]$. Based on the ETBD dataset, we define a new dataset as follows.

$$\mathrm{etbd}' = (\mathrm{etbdID}, \mathrm{latitude}, \mathrm{longitude}, \mathrm{etbdAddress}, \mathrm{ETBD\_heat}). \tag{10}$$

Generally speaking, the grade of each ETBD, ETBD_heat, remains stable for relatively long periods of time, in the order of months. In order to mitigate data deviation, all ETBD-relevant data are stored in a Hadoop Distributed File System (HDFS) and we analyze them using Pig [29] and Perl [30], which are especially designed for big data processing.

As for traffic, its characteristics differ from a city to another, from a day to another and also within the same day. Indeed, morning and evening rush hours may be different for week work days and holidays, and may even differ from a city to another and that is due to many factors such as the city location, residents' behavior, and city regulation. To model morning and evening rush hours for weekdays and holidays with an acceptable level of accuracy, the following algorithm is devised based on big data analysis and statistics.

i. Count weekdays and holidays in time τ, denoted respectively by T_{weekday} and T_{holiday}

$$\tau = T_{\mathrm{weekday}} \cup T_{\mathrm{holiday}}. \tag{11}$$

ii. Count the number of vehicle passing records for each hour of the day during weekdays based on $C_{(\mathrm{vpr},\tau)}$

$$C_{(\mathrm{vpr},\mathrm{T_{weekday}},\mathrm{hour})} = \frac{\sum_{\mathrm{hour}=j} (r_i | c\mathrm{Date}_i \in T_{\mathrm{weekday}}, j \in [1, \ldots, 23])}{|T_{\mathrm{weekday}}|}. \tag{12}$$

iii. Count the number of vehicles for each hour of the day during weekdays based on $C_{(\mathrm{vehicle},\tau)}$, i.e., (13), shown at the bottom of the page.

iv. Get the average speed for each hour of the day during weekdays to represent congestion rate using $C_{(\mathrm{speed},\mathrm{T_{weekday}},\mathrm{hour})}$ divided by $C_{(\mathrm{vpr},\mathrm{T_{weekday}},\mathrm{hour})}$. The lower the average speed is, the more congested the traffic is.

$$\mathrm{Average\_speed}_{\mathrm{weekday}} = \frac{C_{(\mathrm{speed},\mathrm{T_{weekday}},\mathrm{hour})}}{C_{(\mathrm{vpr},\mathrm{T_{weekday}},\mathrm{hour})}}. \tag{14}$$

v. The average speed during holidays can be calculated in the same fashion as $\mathrm{Average\_speed}_{\mathrm{holiday}}$

$$C_{(\mathrm{vehicle},\mathrm{T_{weekday}},\mathrm{hour})} = \frac{\sum_{\mathrm{hour}=j} (r_i | c\mathrm{Date}_i \in T_{\mathrm{weekday}}, j \in [1, \ldots, 23], \text{Group by } v\mathrm{License})}{|T_{\mathrm{weekday}}|} \tag{13}$$

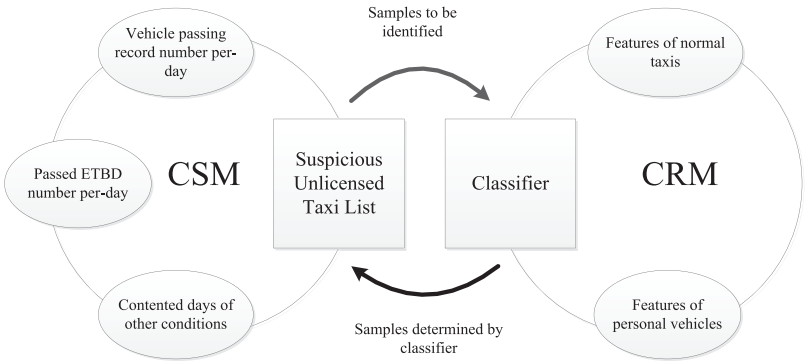

Fig. 1. Unlicensed taxi identification model.

vi. Based on the computed average speed for each hour, we get the morning and evening rush hours for both weekdays and holidays as the periods of time in the morning and evening, respectively, when the average speed takes its lowest values, and that is for both holidays and weekdays.

$$\text{morning\_rush\_hour} = [\text{mrh\_start\_hour}, \text{mrh\_end\_hour}] \quad (15)$$

$$\text{evening\_rush\_hour} = [\text{erh\_start\_hour}, \text{erh\_end\_hour}] \quad (16)$$

mrh_start_hour and mrh_end_hour are the start and end time for moring_hush_hour, and the same meaning of erh_start_hour and erh_end_hour.

IV. Unlicensed Taxi Identification Model

In this section, we will describe in details our proposed Unlicensed Taxi Identification Model (UTIM) based on statistics [31] and machine learning [32]. UTIM consists of two sub-model components, namely *i*) Candidate Selection Model (CSM) and *ii*) Candidate Refined Model (CRM). The former one is based on statistics and the latter is based on machine learning, as depicted in Fig. 1.

CSM screens out a coarse-grained suspected unlicensed taxi candidate list according to vehicles' features, such as vehicle type, vehicle activity level during a day, traversed ETBDs, and days showing similar behavior which can directly get rid of the normal vehicles. This candidate list is subsequently used as input for CRM to get a fine-grained suspected unlicensed taxi candidate list. This model is trained based on the features of the confirmed unlicensed taxis and normal vehicles.

A. Candidate Selection Model (CSM)

Generally, unlicensed taxis keep running in the city for long time periods. They consequently pass high number of ETBDs, resulting in high vehicle passing records. Their vehicle trajectories are usually different from those of normal vehicles, which tend to be predictable [36]–[38]. To reduce false alarms, unlicensed taxis and normal vehicles should be distinguished based on statistics. Obviously, police cars, normal taxis, trucks and vehicles with specific roles can be easily deemed as not unlicensed taxis according to rules of the registered license plate. They can be therefore excluded from the operations of CSM. The CSM model is schematically portrayed in Fig. 2.

The CSM algorithm operates as follows.

a. During a time period τ, the system develops a vehicle list $V = \{v_1, v_2, \ldots, v_n\}$ and the number of vehicle passing records of each day $C_{(\text{vpr,vLicense,day}_\text{i})}$ according to the aforementioned analysis;

$$C_{(\text{vpr,vLicense,day}_\text{i})} = \sum (r_i | c\text{Date}_i \in \text{day}_i, v\text{License}_i = v\text{License}, \text{day}_i \in \tau). \quad (17)$$

b. The system also computes the number of ETBDs traversed during each day $C_{(etbd,vehicle,day_i)}$

$$C_{(\text{etbd,vehicle,day}_\text{i})} = \sum (r_i | c\text{Date}_i \in \text{day}_i, v\text{License}_i = v\text{License}, \text{Group by etbd}). \quad (18)$$

c. The system then filters out specific vehicle types (e.g., police car) according to the rules of license plate from the vehicle list V.

d. $C_{(\text{vpr,vLicense,day}_\text{i})}$ should be greater than the threshold thre_vpr

$$C_{(\text{vpr,vLicense,day}_\text{i})} > \text{thre\_vpr}. \quad (19)$$

e. $C_{(\text{etbd,vehicle,day}_\text{i})}$ should be greater than the threshold thre_etbd

$$C_{(\text{etbd,vehicle,day}_\text{i})} > \text{thre\_etbd}. \quad (20)$$

f. Through the above steps, the system can sort out a vehicle list for each day V_{day_i}.

g. The system then counts the total number of times each vehicle appears in V_{day_i}

$$C_{(v\text{License})} = \sum_{j \in V_{\text{day}_i}} v_i | (v\text{License} \in j, \text{day}_i \in \tau). \quad (21)$$

h. $C_{(v\text{License})}$ should be greater than the threshold thre_day

$$C_{(v\text{License})} > \text{thre\_day}. \quad (22)$$

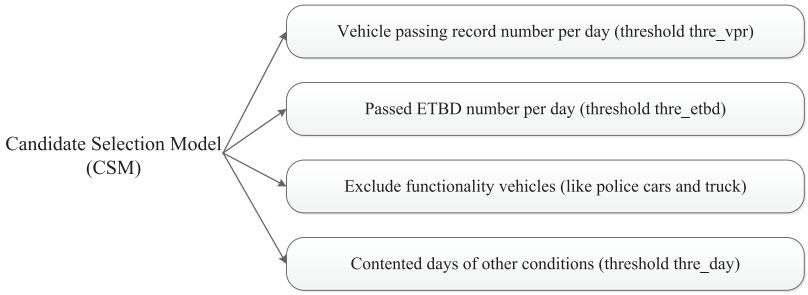

Fig. 2. Candidate selection model.

i. Finally, based on the conditions of inequalities (19), (20), and (22), the system sorts out a coarse-grained list of suspected unlicensed taxi candidates $V' = \{v'_1, v'_2, \ldots, v'_n\}$.

B. Candidate Refined Model (CRM)

After the creation of the coarse-grained unlicensed taxi candidate list V', CRM comes into play and that is to further enhance the accuracy of the proposed scheme in unlicensed taxi identification. In CRM, the unlicensed taxi identification is transformed into a classification issue using data mining. For each vehicle, there are only two states; unlicensed taxi or normal vehicle.

1) Sample Selection: CRM is based on machine learning. For that purpose, we use both positive samples and negative samples which are inputs for model training. As normal taxis function similarly to unlicensed taxis, in regard of their time and spatial characteristics, they are used as positive samples of unlicensed taxis. The samples can be described as:

$$S_{ut} = \{v_1, v_2, \ldots, v_m\}. \tag{23}$$

As for the selection of negative samples, according to feedback and experience from police, most of unlicensed taxis are entry-level vehicles which -are much cheaper. So, the high-end vehicles which are expensive are used as negative samples as:

$$S_{nv} = \{nv_1, nv_2, \ldots, nv_k\}. \tag{24}$$

2) Feature Extraction: Different features affect the accuracy of the proposed unlicensed taxi identification model. In this paper, features based on time and spatial characteristics are considered. This aims for distinguishing unlicensed taxis from normal vehicles to the maximum extent. Based on our experiments, the considered features are as follows.

i. Number of vehicle passing records averaged over predetermined time periods.

The average vehicle passing records is counted for different time periods as show in Tables III and IV, distinguishing between weekdays and holidays. The time periods are determined according to residents' daily activity times, morning rush hour, and evening rush hours. The feature corresponding to the number of vehicle passing records is then created based on the values obtained during each time period as follows:

$$F_1 = \{f_1, f_2, \ldots, f_{18}\}. \tag{25}$$

TABLE III
WEEKDAY TIME PERIODS

Index	Time Period
p_1	$00:00 \sim mrh\_start\_hour_w$
p_2	$mrh\_start\_hour_w \sim mrh\_end\_hour_w$
p_3	$mrh\_end\_hour_w \sim mrh\_end\_hour_w + 1$
p_4	$mrh\_end\_hour_w + 1 \sim 12:00$
p_5	$12:00 \sim 13:00$
p_6	$13:00 \sim erh\_start\_hour_w - 1$
p_7	$erh\_start\_hour_w - 1 \sim erh\_start\_hour_w$
p_8	$erh\_start\_hour_w \sim erh\_end\_hour_w$
p_9	$erh\_end\_hour_w \sim 24:00$

TABLE IV
HOLIDAY TIME PERIODS

Index	Time Period
p_{10}	$00:00 \sim mrh\_start\_hour_h$
p_{11}	$mrh\_start\_hour_h \sim mrh\_end\_hour_h$
p_{12}	$mrh\_end\_hour_h \sim mrh\_end\_hour_h + 1$
p_{13}	$mrh\_end\_hour_h + 1 \sim 12:00$
p_{14}	$12:00 \sim 13:00$
p_{15}	$13:00 \sim erh\_start\_hour_h - 1$
p_{16}	$erh\_start\_hour_h - 1 \sim erh\_start\_hour_h$
p_{17}	$erh\_start\_hour_h \sim erh\_end\_hour_h$
p_{18}	$erh\_end\_hour_h \sim 24:00$

ii. Average vehicle passing records for each ETBD grade

As explained earlier, ETBDs have different grades. Unlicensed taxis and normal vehicles traverse different numbers of ETBDs. The former always run for longer times in the city, so their ETBD distribution is dispersed. Whereas, the latter travel within a small area and their ETBD distribution remains stable. The feature of average vehicle passing records for different ETBD_heat is therefore worth considering. There are h grades in ETBD_heat, so there are $18 * h$ features in this dimension.

$$F_2 = \{f_{19}, f_{20}, \ldots, f_{18*h+18}\}. \tag{26}$$

iii. Number of passed ETBDs averaged over predetermined time periods.

Unlicensed taxis and normal vehicles take different trajectories in different time periods. The average number of ETBDs they traverse is also different. With regard to the average number of passed ETBDs over predetermined time periods, there are $18 * h$ features as follows:

$$F_3 = \{f_{18*h+19}, f_{18*h+20}, \ldots, f_{2*18*h+18}\}. \tag{27}$$

iv. Average vehicle passing records of every day.

Average vehicle passing records of every day is computed for all samples and the features look like:

$$F_4 = \{f_{2*18*h+19}, f_{2*18*h+20}\}. \tag{28}$$

As a result, a total of $(2 * 18 * h + 20)$ features are obtained. For instance, in case ETBDs are classified into three grades, there should be 128 features to train the proposed model. The features, $F = \{f_1, f_2, \ldots, f_{2*18*h+20}\}$, can be further refined by discretizing their values as follows.

i. For a given feature f_i, there is a value set for all samples as below.

$$V_{f_i} = \{\text{value}_1, \text{value}_2, \ldots, \text{value}_{m+k}\}. \tag{29}$$

ii. Cluster V_{f_i} to create a p categories centroid set like $\text{C} = \{c_1, c_2, \ldots, c_p\}$.

iii. Refine the feature f_i into $p+1$ features based on $\text{C} = \{c_1, c_2, \ldots, c_p\}$ and the new feature set becomes $F_{f_i} = \{f'_1, f'_2, \ldots, f'_{p+1}\}$ which includes $p+1$ features. The new feature value $\text{V}(f'_i)$ is refined as below.

$$\text{V}(f'_i) = \begin{cases} 1; \text{i f} C_{i-1} < \text{value}_i \le C_i, i > 1 \\ 1; \text{i f} 0 \le \text{value}_i \le C_i, i = 1 \\ 0; \text{i f value}_i > C_i, i \ge 1. \end{cases} \tag{30}$$

iv. For all features, $F = \{f_1, f_2, \ldots, f_{2*18*h+20}\}$, repeat the above steps till the new refined feature set $F' = \{f'_1, f'_2, \ldots, f'_q\}$ is obtained, whereby

$$\text{q} = \sum_{i=1}^{2*18*h+20} C(\text{Centroid}_i) \tag{31}$$

Centroid_i denotes the centroid set by clustering V_{f_i} and $C(\text{Centroid}_i)$ denotes its size.

3) Model Training: The proposed model is trained based on SVM to identify whether a vehicle is an unlicensed taxi or a normal vehicle. The basic idea is to map the low-dimensional space curve (or surface) to a high-dimensional space line or plane. After this step, the feature is linearly separable in high-dimensional space. It is trained by the cross-training model sample, as shown below.

i. Normalize the feature set $F' = \{f'_1, f'_2, \ldots, f'_q\}$ to remove the features-the value of which is zero and mark the feature sequence;

ii. Select the Radial Basis Function as the kernel function to train the model

$$\text{k}(\|\text{x} - \text{xc}\|) = \exp\left\{-\frac{\|x - xc\|^2}{2\sigma^2}\right\} \tag{32}$$

whereby xc denotes the kernel function center, and σ denotes the function's width parameter;

iii. Train the optimal model parameters ε and σ, which are the penalty factor and nuclear parameter. The two parameters enable the model to have a high precision for the model training.

TABLE V
COMPUTING PLATFORM SETUP

Index	Item	Big data analysis server
Hardware	CPU	2 processors, Xeon 6, E5-2630 2.3GHz
	Cache	20MB each one
	Memory	64GB(4X16GB)
Software	System	Red Hat Enterprise Linux
	HBase	hbase-0.94.14-security
	Hadoop	hadoop-1.2.1
Server number		20
Nodes number		80

TABLE VI
ETBD HEAT CLASSIFICATION

ETBD heat	$C_{(etbdID,1\ month)}$ *(million)*	*Number of ETBDs*
$ETBD\_heat_1$	> 3.8	8
$ETBD\_heat_2$	> 1.5	66
$ETBD\_heat_3$	< 1.5	229

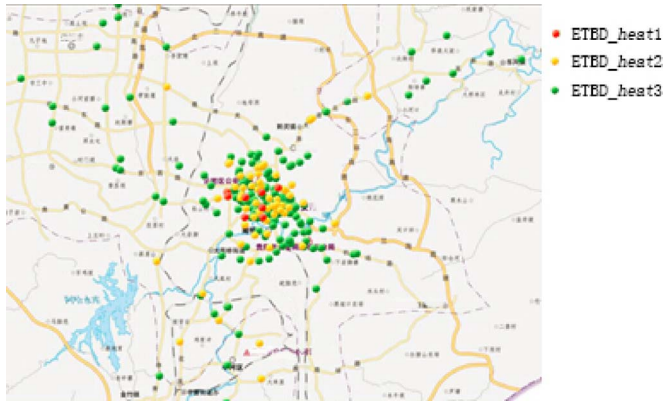

Fig. 3. ETBD classification on map.

V. RESULTS AND ANALYSIS

Having described in details our proposed model, we now concentrate on its evaluation. For this purpose, we use a dataset consisting of vehicle passing records generated by 1,193,281 vehicles over a period of 31 days. The number of vehicle passing records is 340,679,449, the size of which in HDFS is about 28 G. The records are made by 303 ETBDs installed along the main roads in the city. On average, each ETBD can generate 36,270 vehicle passing records every day. To train our proposed model, we use 6,868 normal taxis and 3,760 personal vehicles as positive samples and negative samples, respectively. As a computing platform setup, we use 20 physical servers which are virtualized to 80 virtual nodes. They are built in Hadoop and HBase. Table V shows other details about the computing platform setup.

As explained earlier, vehicle passing records $C_{(\text{vpr,etbID,1 month})}$, generated in one month, are used for ETBD classification. By clustering the data, we obtain three grades of ETBD for the model training. The clustering results are shown in Table VI and the corresponding locations on the map are illustrated in Fig. 3. As shown in Table VI, for the first ETBD grade (ETBD_heat_1), the number of vehicle

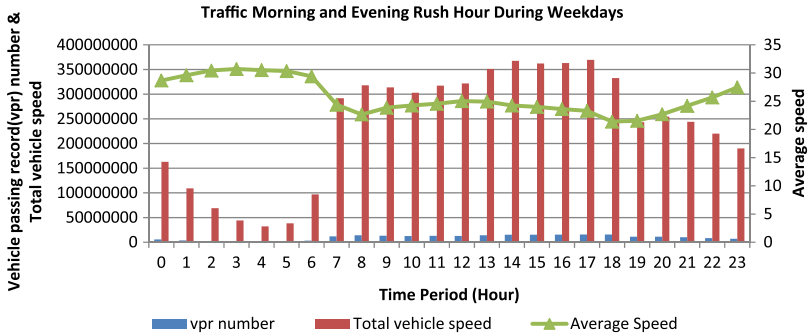

Fig. 4. Average speed in weekdays.

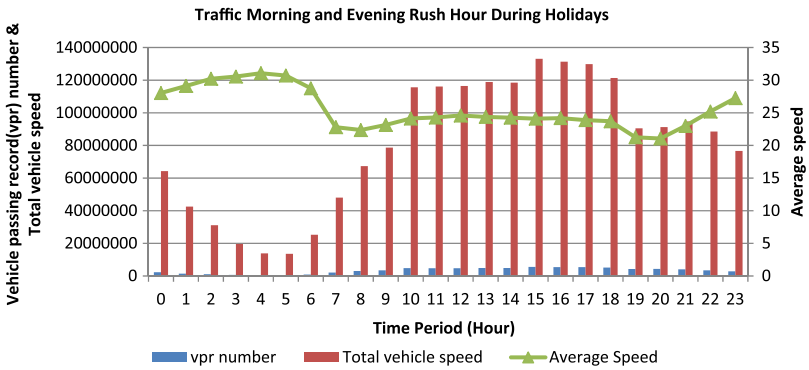

Fig. 5. Average speed in holidays.

TABLE VII
CSM EVALUATION RESULTS

index	*thre_vpr*	*thre_etbd*	*thre_day*	Candidate number	Retrieved number of vut	vut	accuracy ratio
1	70	30	15	4,532	56	136	41.18%
2	60	30	15	6,126	103	136	75.74%
3	50	25	15	8,256	117	136	86.03%
4	**50**	**20**	**10**	**11,231**	**135**	**136**	**99.26%**
5	50	15	10	32,416	136	136	100%
6	30	20	10	56,178	136	136	100%

passing records exceeds 3.8 million and it has 8 ETBDs. As for ETBD_heat$_2$, the number of vehicle passing records is between 1.5 million and 3.8 million. 66 ETBDs belong to this grade; for the last grade ETBD (ETBD_heat$_3$), the number of vehicle passing records is less than 1.5 million and there are 229 relevant ETBDs. Fig. 3 shows the geographical distribution of ETBDs on the map. Different grades are highlighted using different colors. On the map, there are eight red points, 66 yellow points and 229 green points, referring to ETBD_heat$_1$, ETBD_heat$_2$ and ETBD_heat$_3$, respectively.

To compute the city traffic morning and evening rush hours for weekdays and holidays, one month vehicle passing record $C_{(\text{vpr,etbID,1 month})}$ is used. The results are shown in Figs. 4 and 5. Indeed, Fig. 4 shows the congestion rate during weekdays. "vpr" refers to vehicle passing record. For the figure, it can be deduced that the morning rush hour is from 6:00 to 8:00 and the evening rush hour is from 17:00 to 20:00. Fig. 5 shows the congestion rate during holidays. The morning rush hour is also from 6:00 to 8:00 and the evening rush hour is from 18:00 to 21:00.

TABLE VIII
WEEKDAY AND HOLIDAYS TIME PERIODS

Index	Weekday Time Periods	Index	Holiday Time Periods
p_1	00:00 ~ 6:00	p_{10}	00:00 ~ 6:00
p_2	6:00 ~ 8:00	p_{11}	6:00 ~ 8:00
p_3	8:00 ~ 9:00	p_{12}	8:00 ~ 9:00
p_4	9:00 ~ 12:00	p_{13}	9:00 ~ 12:00
p_5	12:00 ~ 13:00	p_{14}	12:00 ~ 13:00
p_6	13:00 ~16:00	p_{15}	13:00 ~17:00
p_7	16:00 ~ 17:00	p_{16}	17:00 ~ 18:00
p_8	17:00 ~ 20:00	p_{17}	18:00 ~ 21:00
p_9	20:00 ~ 24:00	p_{18}	21:00 ~ 24:00

To evaluate the proposed UTIM model, we first evaluate its CSM component. As UTIM is based on machine learning, the more real unlicensed taxis it identifies, the better CSM becomes. For this purpose, we use a preliminary list of unlicensed taxis already verified by the local police department. CSM can tell how many real verified unlicensed taxis are identified in the candidate list using different model parameters. CSM is evaluated based on the accuracy ratio of the number of identified unlicensed taxis to the total number of unlicensed

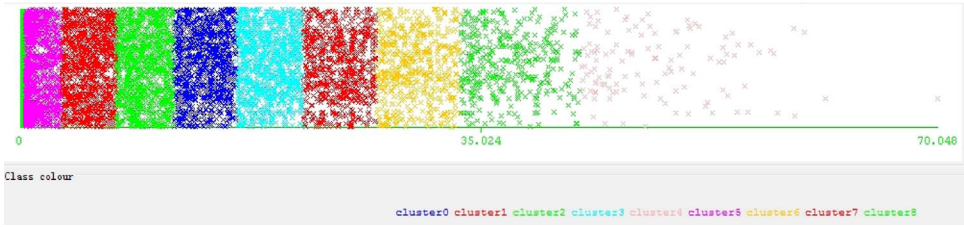

Fig. 6. Feature refinement results.

taxis as verified by the local police department. Intuitively, thre_vpr, thre_etbd and thre_day are three key parameters that could impact the model quality.

Table VII shows the CSM evaluation results. "vut" denotes the total number of unlicensed taxis verified by the local police department. "Candidate number" indicates the total number of vehicles identified as unlicensed taxis by CSM. "Retrieved number of vut" indicates how many vehicles in the candidate list are indeed unlicensed taxis as verified by the police department. The accuracy ratio is the ratio of the number of vut identified by CSM to vut. Table VII shows that CSM functions best when the parameters thre_vpr, thre_etbd and thre_day are set to 50, 20 and 10, respectively, as it generates with high accuracy a list of only 11231 candidates that can be easily processed by CRM.

To evaluate the performance of CRM, 6,868 positive samples and 3,760 negative samples are used (i.e., $m = 6,868$ in S_{ut} and $k = 3,760$ in S_{nv}). Furthermore, according to the weekday/holiday morning/evening rush hours determined from Figs. 4 and 5, the weekday/holiday time periods can be divided as shown in Table VIII. Additionally, as ETBDs are classified into three grades, the envisioned samples have 128 features—$F = \{f_1, f_2, \ldots, f_{128}\}$. Consequently, in the feature refinement process, there should be 10628 sample values for each feature f_i, $V_{f_i} = \{\text{value}_1, \text{value}_2, \ldots, \text{value}_{10628}\}$.

Fig. 6 shows the feature refinement results. Different color sections represent different feature groups. After feature refinement, the new feature set becomes—$F' = \{f'_1, f'_2, \ldots, f'_{1280}\}$ consisting of 1,280 features. The value of each feature and for each sample can be collected. Then, the final feature values form a matrix which has $1,280 * 10,628$ elements that are used to train CRM. Fig. 7 shows the results of the CRM model training. The model accuracy for all samples is 98.6029%. The parameter ε is 8 and σ is 0.0078125. The candidate list V' is used as input for CRM to identify the fine-grained suspected unlicensed taxis. Table IX shows the CRM results, indicating the number of unlicensed taxis (out of CSM's candidate list) that are identified by CRM.

To validate the CRM results, trajectories of all unlicensed taxis identified by CRM were carefully checked. Figs. 8 and 9 show the example of a vehicle xxx0C38 among those identified by CRM. The vehicle kept running for long times in the city. It carried different passengers in short time as shown in the two figures. For instance, Fig. 8 shows the vehicle xxx0C38 carrying a female passenger wearing a red dress at 16:28:41 on Jan 22nd in 2014. Fig. 9 shows the vehicle carrying another female passenger wearing a grey dress 20 minutes later, at 16:48:13. During the same day, such situation happened several

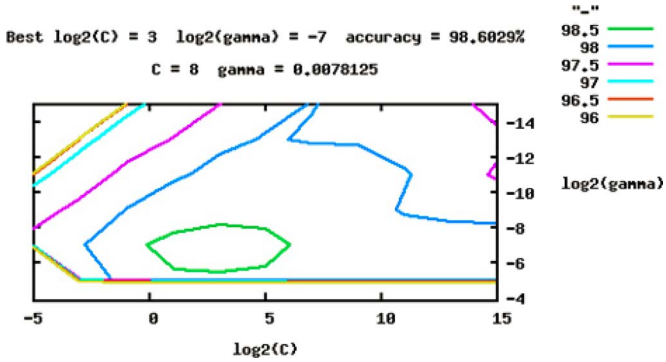

Fig. 7. CRM training results.

TABLE IX
CRM RESULTS

ε	σ	Candidate number	Retrieved unlicensed taxi number
8	0.0078125	11,231	4,348

Fig. 8. Vehicle xxx0C38 carrying a female passenger wearing a red dress.

Fig. 9. Vehicle xxx0C38 carrying a female passenger wearing a grey dress.

times, each time with a new passenger on board, confirming that this vehicle is indeed an unlicensed taxi. All 4,348 vehicles identified by CRM as unlicensed taxis were carefully investigated in the same manner along with the professional police officer and finally 3,768 unlicensed taxis were indeed confirmed, yielding an overall accuracy of 86.667%.

VI. Conclusion

As an important contribution to D^{2}ITS, this paper described an efficient model for the identification of unlicensed taxis. The efficiency of the model in accurately identifying unlicensed taxis is demonstrated through real-life implementation of the model and its application to real-life traffic in a medium size Chinese city. Indeed, using the proposed model, unlicensed taxis were identified from over 800,000 vehicles at an accuracy equal to 86.667%, a value high enough to enable relevant legal institutions (e.g., police department and public transportation department) to efficiently mitigate the phenomenon of unlicensed taxis in China and other countries suffering from the same issue. Whilst the obtained results are encouraging, the learning phase of the model can be further enhanced. Furthermore, admittedly, the characteristics used to distinguish unlicensed taxis from normal ones may not be sufficient. Other characteristics are yet to be defined and considered. This defines one of the future research directions of the authors in this particular area of research.

References

[1] Y. Wang, W. Heng, and Q. Shi, "Current status for ITS national architecture development," *J. Intell. Transp. Syst.*, vol. 9, no. 3, pp. 9–16, 2001.

[2] K. Yin, "How data mining is applied in intelligent transportation field," *Inf. Commun.*, no. 10, pp. 92–93, 2013.

[3] J. Zhang *et al.*, "Data-driven intelligent transportation systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1624–1639, Dec. 2011.

[4] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Netw. Appl.*, vol. 19, no. 2, pp. 171–209, 2014.

[5] N. Shi, "Intelligent Detection of License Plate-Related Illegal vehicles Using Bayonet Images," Ph.D. dissertation, Nanjing Normal Univ., Nanjing, China, 2013.

[6] Y. Zhang and Y. Zhang, "Taxi management mode and unlicensed taxi problem in China," *Prod. Investigation*, no. 9, pp. 157–159, 2011.

[7] H. Hu, Y. Wen, T. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 2, pp. 652–687, Jul. 2014.

[8] C. Hsu, C. Chang, and C. Lin, "A Practical Guide to Support Vector Classification, 2003. [Online]. Available: https://www.cs.sfu.ca/people/Faculty/teaching/726/spring11/svmguide.pdf

[9] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, Jan. 2008.

[10] T. White, *Hadoop: The Definitive Guide*. Sebastopol, CA, USA: O'Reilly Media, 2012.

[11] Y. Malinovskiy, Y. Wu, and Y. Wang, "Video-based monitoring of pedestrian movements at signalized intersections," *Transp. Res. Rec.: J. Transp. Res. Board*, vol. 2073, pp. 11–17, 2008.

[12] Y. Malinovskiy, J. Zheng, and Y. Wang, "Model-free video detection and tracking of pedestrians and bicyclists," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 24, no. 3, pp. 157–68, Apr. 2009.

[13] Y. Malinovskiy, Y. Wu, and Y. Wang, "Video-based vehicle detection and tracking using spatiotemporal maps," *Transp. Res. Rec.: J. Transp. Res. Board*, vol. 2121, pp. 81–89, 2009.

[14] X. Ma, Y. Jan, and Y. Wang, "DRIVE Net: An e-science of transportation platform for data sharing, visualization, modeling, and analysis," *Transp. Res. Rec.: J. Transp. Res. Board*, vol. 2215, pp. 37–49, 2011.

[15] X. Ma, E. McCormack, and Y. Wang, "Processing commercial GPS data to develop a web-based truck performance measures program," *Transp. Res. Rec.: J. Transp. Res. Board*, vol. 2246, pp. 92–100, 2011.

[16] X. Ma, Y. Wang, F. Chen, and J. Liu, "Transit smart card data mining for passenger origin information extraction," *J. Zhejiang Univ. Sci. C*, vol. 13, no. 10, pp. 750–760, Oct. 2012.

[17] Y. Malinovskiy, N. Saunier, and Y. Wang, "Pedestrian travel analysis using static bluetooth sensors," *Transp. Res. Rec.: J. Transp. Res. Board*, vol. 2299, pp. 137–149, 2012.

[18] X. Gong and X. Liu, "A data-mining-based algorithm for traffic network flow forecasting," in *Proc. Int. Conf. Integr. Knowl. Intensive Multi-Agent Syst.*, 2003, pp. 243–248.

[19] J. W. C. Van Lint, "Online learning solutions for freeway travel time prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 9, no. 1, pp. 38–47, Mar. 2008.

[20] H. Jula, M. Dessouky, and P. A. Ioannou, "Real-time estimation of travel times along the arcs and arrival times at the nodes of dynamic stochastic networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 9, no. 1, pp. 97–110, Mar. 2008.

[21] C. Bi, X. Ma, Y. Zhang, and Y. Wang, "Developing a cluster-based algorithm for collision hotspot identification," in *Proc. CICTP*, Aug. 2014, pp. 2381–2395.

[22] Y. Jin and J. Wang, "The way to reduce the vehicles using the copied licenses based on traffic information collection," *Chin. Water Transp.*, no. 9, pp. 173–176, 2007.

[23] E. Murakami and D. P. Wagner, "Can using Global Positioning System(GPS) improve trip reporting," *Transp. Res. Part C, Emerging Technol.*, vol. 7, no. 2/3, pp. 149–165, Apr.–Jun. 1999.

[24] Y. Lou *et al.*, "Map-matching for low-sampling-rate GPS trajectories," in *Proc. ACM GIS*, 2009, pp. 352–361.

[25] J. Shang, Y. Zheng, W. Tong, E. Chang, and Y. Yu, "Inferring gas consumption and pollution emissions of vehicles throughout a city," in *Proc. ACM KDD*, 2014, pp. 1027–1036.

[26] C. Bi, W. Yuan, B. Yan, P. Deng, and F. Chen, "An algorithm to detect the automobiles using the copied vehicle license," in *Proc. CCISA*, 2014, pp. 225–231.

[27] Q. Wang *et al.*, "A new distributed strategy to schedule computing resource," in *Proc. CCISA*, 2014, pp. 216–224.

[28] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Prob.*, 1967, pp. 281–297.

[29] Hadoop: Apache Pig, Accessed Aug. 25, 2014. [Online]. Available: http://pig.apache.org/

[30] Perl, Accessed Aug. 25, 2014. [Online]. Available: http://www.perl.org/

[31] Z. Ivezic, A. J. Connolly, J. T. VanderPlas, and A. Gray, " *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data*. Princeton, NJ, USA: Princeton Univ. Press, 2014.

[32] S. Sharma, J. Agrawal, S. Agarwal, and S. Sharma, "Machine learning techniques for data mining: A survey," in *Proc. ICCIC*, 2013, pp. 1–6.

[33] L. Foschini, T. Taleb, A. Corradi, and D. Bottazzi, "M2M-based metropolitan platform for IMS-enabled road traffic management in IoT," *IEEE Commun. Mag.*, vol. 49, no. 11, pp. 50–57, Nov. 2011.

[34] T. Taleb, A. Benslimane, and K. Ben Letaief, "Towards an effective risk-conscious and collaborative vehicular collision avoidance system," *IEEE Trans. Veh. Technol.*, vol. 59, no. 3, pp. 1474–1486, Mar. 2010.

[35] T. Taleb and K. Ben Letaief, "A cooperative diversity based handoff management scheme," *IEEE Trans. Wireless Commun.*, vol. 9, no. 4, pp. 1462–1471, Apr. 2010.

[36] A. Benslimane, T. Taleb, and R. Sivaraj, "Dynamic clustering-based adaptive mobile gateway management in integrated VANET-3G heterogeneous wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 3, pp. 559–570, Mar. 2011.

[37] A. Nadembega, A. Hafid, and T. Taleb, "Mobility prediction-aware bandwidth reservation scheme for mobile networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 6, pp. 2561–2576, Jun. 2014.

[38] A. Nadembega, A. Hafid, and T. Taleb, "An integrated predictive mobile-oriented bandwidth-reservation framework to support mobile multimedia streaming," *IEEE Trans. Wireless Commun.*, vol. 13, no. 12, pp. 6863–6875, Dec. 2014.

[39] A. Nadembega, A. Hafid, and T. Taleb, "A destination & mobility path prediction scheme for mobile networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 6, pp. 2577–2590, Jun. 2014.

[40] T. Taleb and A. Ksentini, "VECOS: A vehicular connection steering protocol," *IEEE Trans. Veh. Technol.*, vol. 64, no. 3, pp. 1171–1187, Mar. 2014.

[41] T. Taleb, K. Samdanis, and A. Ksentini, "Supporting highly mobile users in cost-effective decentralized mobile operator networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 7, pp. 3381–3396, Sep. 2014.

[42] Unlicensed Taxi Endangerment, Accessed Jun. 10, 2015. [Online]. Available: http://yanzhao.yzdsb.com.cn/system/2014/08/29/013886729.shtml

[43] Unlicensed Taxi Status, Accessed Jun. 10, 2015. [Online]. Available: http://theory.people.com.cn/GB/40537/9799567.html

[44] J. Wan *et al.*, "VCMIA: A novel architecture for integrating vehicular cyber-physical systems and mobile cloud computing," *Mobile Netw. Appl.*, vol. 19, no. 2, pp. 153–160, 2014.

[45] J. Wan, D. Zhang, S. Zhao, L. T. Yang, and J. Lloret, "Context-aware vehicular cyber-physical systems with cloud support: Architecture, challenges and solutions," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 106–113, Aug. 2014.

[46] J. Liu *et al.*, "A survey on position-based routing for vehicular ad hoc networks," *Telecommun. Syst.*, 2015. [Online]. Available: 10.1007/s11235-015-9979-7

[47] F. Chen *et al.*, "Data mining for the internet of things: Literature review and challenges," *Int. J. Distrib. Sensor Netw.*, vol. 2015, no. 9, 2015, Art. ID 431047.

Wei Yuan received the B.S. degree from Wuhan University, Wuhan, China, in 2011. He is currently working toward the Ph.D. degree in the Parallel Computing Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing, China. He has authored/coauthored four academic papers. His research interests include information retrieval, sponsored search, data mining, parallel computing, smart city, big data, and internet system.

Pan Deng received the Ph.D. degree in computer software and theory from Beijing University of Aeronautics and Astronautics, Beijing, China, in 2011. She is currently an Associate Professor with the Parallel Computing Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing, where she is also the Executive Deputy Director. She is also currently the Executive Vice President of Guiyang Academy of Information Technology, Guiyang, China, and the Deputy Director of Guiyang Technology Bureau, Guiyang. She has directed over 19 science projects as a project leader and has authored/coauthored over 25 scientific papers, four Chinese patents, and 16 software copyrights. Her research interests include smart city, big data, cloud computing, parallel computing, and Internet of things.

Tarik Taleb (SM'10) received the B.E. degree (with distinction) in information engineering and the M.Sc. and Ph.D. degrees in information sciences from Tohoku University, Sendai, Japan, in 2001, 2003, and 2005, respectively. He is currently a Professor with the School of Electrical Engineering, Aalto University, Espoo, Finland. Prior to his current academic position, he was a Senior Researcher and a 3GPP Standards Expert at NEC Europe Ltd. He was then leading the NEC Europe Laboratories Team working on R&D projects on carrier cloud platforms. Before joining NEC and until March 2009, he was an Assistant Professor with the Graduate School of Information Sciences, Tohoku University, in a laboratory fully funded by KDDI. He has been also directly engaged in the development and standardization of the evolved packet system as a member of 3GPP's System Architecture working group. His research interests lie in the field of mobile core, mobile cloud networking, network function virtualization, software-defined networking, mobile multimedia streaming, and social media networking. Prof. Taleb is an IEEE ComSoc Distinguished Lecturer. He is serving as a Chair of the Wireless Communications Technical Committee. He is/was on the editorial board of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE Wireless Communications Magazine, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE Communications Surveys & Tutorials, and a number of Wiley journals. He has been a recipient of numerous awards, including the prestigious IEEE ComSoc Asia-Pacific Best Young Researcher award and the TELECOM System Technology Award from the Telecommunications Advancement Foundation. Some of his research studies have been also awarded best paper awards at prestigious conferences.

Jiafu Wan (M'11) received the Ph.D. degree in mechatronic engineering from South China University of Technology (SCUT), Guangzhou, China, in June 2008. He is currently an Associate Professor at SCUT, where he held a postdoctorate position in Computer Science and Engineering from October 2008 to June 2012. He is a project leader of several projects (e.g., NSFC). He has authored/coauthored one book and over 80 scientific papers (with over 30 indexed by ISI SCIE and over 40 indexed by EI Compendex) and has been cited over 1200 times. His research interests include wireless sensor networks, cyberphysical systems, Internet of things, mobile cloud computing, and embedded systems. Dr. Wan is a Senior Member of CCF and a member of ACM. He was a Workshop Chair of M2MC2012, M2MC2013, and MCC2013. He is a Managing Editor of IJAACS (EI) and IJART (EI).

Chaofan Bi received the master's degree from the University of Washington, Seattle, WA, USA. He is currently an Assistant Research Associate with the Parallel Computing Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing, China. He has authored/coauthored seven academic papers. His research interests focus on traffic safety, transportation big data, and transit performance measurement.